Author:
Leung, Gloria T Y

Title:
Evaluation of the Validity of Large-scale Examinations

EVALUATION OF THE VALIDITY OF LARGE-SCALE EXAMINATIONS

Gloria Tsz Yim Leung

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Education in the Faculty of Social Science and Law

(Word Count: 43 215)

ABSTRACT

In view of the inadequacy in the literature on the methodological aspects of assessment validation, this dissertation aims to devise a validation process for large-scale examinations. To illustrate the implementation and the impact of the process, an evaluation of the validity of the 2015 Hong Kong Diploma of Secondary Education (HKDSE) Liberal Studies (LS) Examination was conducted.

Based on Messick's (1995) definition of assessment validity and Kane's (2013, 2015) Argument-based Approach, the content and substantive validity of the 2015 LS Examination has been evaluated by drawing quantitative and qualitative evidence from multiple sources (both primary and secondary): a live script study, nominal group discussions among examiners and a think-aloud study. Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the New Taxonomy (Marzano et al., 2008) and Kuhn's (2001, 2005) KPI model have been deployed as the analytical framework to investigate the differentiation of candidates' performance and the assessment of higher-order thinking skills by the LS Examination.

The dissertation demonstrates how test developers can gather multi-faceted evidence for a large-scale examination and how they may be informed of aspects that require further improvement. In the evaluation of the LS Examination, it is evident that (i) the examination differentiated candidates across five Levels of Performance by the skill domains stipulated on the Level Descriptors (with the exceptions of *Evaluation* and the consideration of *Cultures/Values* between the lowest 2 levels of achievement (Levels 1 and 2)); (ii) the differentiation of performance complies with cognitive models; and (iii) higher-order thinking skills, including high-order *Information-handling* skills, *Dispositions* of relevant knowledge and concepts, *Argument Formulation* by integrating evidence *and Meta-level* skills, were demonstrated by candidates in the examination.

The validation process has been evaluated in terms of the methodology, the transferability to the evaluation of examination processes, the implications for test development, as well as the limitations to be overcome in future evaluation studies of public examinations.

DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGHNED: _____ DATE: 2 January, 2020_____

DEDICATION

A dedication… a dedication…

to the one I love…

Roger

ACKNOWLEDGEMENTS

Words cannot express my gratitude to Professor William Browne for his valuable time and advice throughout these years. Had it not been with his insightful and expert advice, especially on the quantitative analysis which I found most difficult, I should not have been able to finish my dissertation. I always wonder how he could squeeze time out of his busy schedule to provide me with very detailed comments and to respond to my questions so efficiently.

I would also like to extend my gratitude to Dr Janet Orchard for her effort in helping me to search for such a wonderful Supervisor. I was at a time very frustrated for being turned down by several professors and having waited for half a year for a Supervisor.

Thanks to Dr Eric Fung for his advice and help in the joint study with the Hong Kong Academy for Gifted Education. Also, thanks to the Hong Kong Academy for Gifted Education and the Hong Kong Examinations and Assessment Authority for granting me the approval for the use the data from the Joint Study and the live scripts of the 2015 HKDSE LS Examination.

Last but not the least, thanks to my seniors, Dr Merry Keung and Mr Ka-yiu Lo for their support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER 1 INTRODUCTION

The validity of educational assessment has been a contentious field among measurement experts since as early as the 1950s. Messick (1995) defined validity as a process for gathering evidence for "inferences about score meaning or interpretation and about the implications for action" (p.5). The historical epi-centre of contentions was around the nature of validity. Various emphases in validity evaluation, including the content, external criterion and construct (as defined in the following), have been advocated by measurement experts. Messick (1995) unified various approaches under the term **construct validity**, which evaluates validity from various aspects, including content and substantive components. The following distinctions between three types of validity will be used:

> **Content validity** *investigates whether the assessment is appropriate in measuring the content domains stipulated in the curriculum or course specifications.*
>
> **Criterion validity** *focuses on how well the assessment results measure the "true value" (Cureton, 1951, as cited in Kane, 2013, p.18) of a criterion (such as an attribute).*
>
> **Construct validity** *focuses on whether the intended construct (such as, the thinking skills) has been assessed (Cronbach & Meehl, 1955, as cited in Newton et al., 2014). Messick (1995) used the term* ***Construct Validity*** *to integrate various approaches of validity, including content,* ***substantive*** *(the validity of the assessment of cognitive processes), generalisability, external and consequential components.*

More recently, the evaluation of validity has emerged as a research focus for assessment theorists and various approaches for validation have been put forward. However, the literature on assessment validity is heavily inclined to the theoretical epistemology, especially the definition of validity. Though DeLuca (2011) postulated a methodology for evaluating assessment validity by dialectic and hermeneutic enquiries, few research studies operationalised the proposed

methodology. Empirical research on the validity of large-scale examinations predominantly involves psychometric modelling. In this regard, this dissertation investigates an avenue of research direction which puts into practice a method for evaluating assessment validity. The dissertation will adopt Kane's (2013, 2015) Argument-based Approach, which is one of the most widely accepted approaches, to investigate the validity of a core subject in the Hong Kong Diploma of Secondary Education (the HKDSE), namely the Liberal Studies (LS) Examination.

This research aims to devise a process for evaluating the content and substantive aspects of construct validity[1] (Messick, 1995), by employing empirical data from the 2015 HKDSE LS written examination as a case study. In addition to the primary data of answer scripts of the 2015 HKDSE LS Examination from the HKEAA website, secondary data were taken from a joint study with the Hong Kong Academy for Gifted Education[2] (HKAGE) in 2015-2016. The research questions to be considered are:

(1) To what extent is the content validity of the 2015 HKDSE examination justified?
(2) To what extent is the substantive validity of the 2015 HKDSE examination justified?
(2a) Can the examination differentiate the Levels of Performance of candidates?
(2b) Can the 2015 HKDSE LS Examination assess the higher-order thinking skills of candidates specified in the Level Descriptors?

The assessment validity process proposed was based on the definition of validation by Messick (1995). The focus on content and substantive validity will be justified in Chapter 2. Along this line of thinking, a procedure based on evidence from multiple sources was devised and put into practice for the evaluation of the appropriateness of the interpretation and use of the results of

[1] *Content and substantive aspects are various criteria for evaluating the validation of assessments as suggested by Messick (1995).*
[2] *The joint study aimed to analyse the performance of candidates of the 2015 HKDSE LS Examination, who were the members of the HKAGE. It was conducted from September 2015 to March 2016.*

some candidates from the 2015 HKDSE LS Written Examination.

Deploying Kane's (2013, 2015) Argument-based Approach as the theoretical framework and guided by a pragmatic research paradigm (Morgan, 2007), quantitative and qualitative evidence was gathered from multiple sources to justify the following Validity Arguments , which state the criteria for evaluating the validity of the examination:

> *(1) The Assessment Objectives and the assessment criteria of the 2015 HKDSE LS Examination align with the Level Descriptors;*
> *(2) The Level Descriptors appropriately differentiate the performance of candidates;*
> *(3) The 2015 LS Examination assesses the higher-order thinking skills of candidates specified in the Level Descriptors.*

A content analysis was conducted to evaluate the alignment among the Assessment Objectives, the assessment criteria and the Level Descriptors (Validity Argument (1)). Live scripts, nominal group discussions and think-aloud protocols, which were from both primary and secondary sources[3] of evidence for assessment validation suggested by Shaw et al. (2012), were analysed quantitatively and qualitatively with reference to the cognitive models of Bloom (1956), Anderson & Krathwohl (2001), Marzano et al. (2008) and Kuhn (2001, 2005) to examine Validity Argument (2) the differentiation of the levels of thinking skills; and (3) the assessment of higher-order thinking skills by the examination.

Content and substantive validity will be evaluated from multiple sources with a view to informing the examination developers of ways for improvement. The factors contributing to the applicability and the limitations of the proposed validation process will be analysed.

The LS examination is chosen for illustrating the validation process because it is a relatively new

[3] *Parts of the scores to the live scripts were primary sources. The others were secondary sources.*

core subject for all Secondary Six students (the last year of senior secondary school[4]) in Hong Kong and research studies are lacking in this area. The first cohort of students of this three-year curriculum took the public examination (the HKDSE) in 2012. In response to the concerns of stakeholders on the validity of this examination, (as suggested in *Continual Renewal from Strength to Strength - Report on the New Academic Structure Medium-term Review* (Curriculum Development Council, Hong Kong Examinations and Assessment Authority (HKEAA) & Education Bureau, 2015)) a rigorous procedure for evaluating this examination is essential. Mason (2007) explicitly pointed to the importance of justifying the validity of an assessment for high-stakes purposes:

> *... "if the purpose is to be the only source of information on which critical life-changing decisions about people who work and learn in the schools are based, then the tests must be shown to be valid for that purpose." (p.44)*

In view of the importance attached to the HKDSE examination, empirical data were drawn from this examination with an aim of illuminating a practicable validation process for a large-scale examination. The curriculum, the assessment framework and the grading mechanism of this subject will be further elaborated in the following section.

1.1 Background – the HKDSE LS Examination

The Hong Kong Diploma of Secondary Education (HKDSE) Liberal Studies (LS) Examination is an assessment of the performance of candidates in a three-year curriculum in senior secondary education. As one of the four core subjects, candidates are required to attain a minimum of Level 2 in the 7-level (Levels 1 to 5, 5* and 5**) grading mechanism in LS for admission to government-funded degree programmes.

[4] *The Senior Secondary School curriculum is offered for students aged between 15 and 18 in all government-funded secondary schools.*

As stipulated in the *Curriculum and Assessment Guide* of LS (the Curriculum Development Council and the HKEAA, 2014), this subject adopts an issue-enquiry approach, rather than content-based approach, aiming to encourage "students to develop a capacity for independent learning in the pursuit of knowledge" (p.4). Students are expected to "develop multiple perspectives on perennial and contemporary issues" and "become independent thinkers" (p.4) with skills for life-long learning. Based on the premise of the curriculum, Assessment Objectives were established, setting forth guidelines for the public examinations of the subject. As detailed in Table 1.1, the assessments of LS should evaluate candidates' ability to apply knowledge and concepts relevant to contemporary issues (Objectives *a*, *b, c and d*); demonstrate high-order thinking skills (for example analysing issues, solving problems, making judgements and conclusions, providing suggestions) (Objective *e, g, h, j, m*); make considerations from multiple perspectives (Objectives *d, f, i and j*); demonstrating enquiry skills[5] (Objective *k*); communicating clearly (Objective *l*) and demonstrating understanding and appreciation of different cultures and universal values (Objectives *n* and *o*).

[5] *Objective k is out of the scope of this dissertation as it is assessed by the School-based Assessment, which is an Independent Enquiry Study, rather than by the written examination.*

Table 1.1: Assessment Objectives of LS (HKEAA, 2017) (The key skills are in bold.)

The objectives of this assessment are to evaluate candidates' abilities:

- *(a) to demonstrate a sound understanding of the key ideas,* ***concepts and terminologies*** *of the subject;*
- *(b) to* ***make conceptual observations*** *from information resulting from enquiry into issues;*
- *(c) to* ***apply relevant knowledge and concepts to contemporary issues****;*
- *(d) to identify* ***and analyse the interconnectedness and interdependence amongst personal, local, national, global and environmental contexts****;*
- *(e) to recognise the influence of personal and social values in analysing contemporary issues of human concern;*
- *(f) to draw critically upon their own experience and their encounters within the community, and with the environment and technology;*
- *(g) to discern views, attitudes and values stated or implied in any given factual information;*
- *(h) to* ***analyse issues*** *(including their moral and social implications),* ***solve problems, make sound judgments and conclusions and provide suggestions, using multiple perspectives****, creativity and appropriate thinking skills;*
- *(i) to interpret information* ***from different perspectives****;*
- *(j) to consider and comment* ***on different viewpoints*** *in their handling of different issues;*
- *(k) to self-manage and reflect upon the implementation of successive stages of* ***the enquiry learning process*** *in terms of time, resources and attainment of the objectives of the enquiry;*
- *(l) to* ***communicate clearly*** *and accurately in a concise, logical, systematic and relevant way;*
- *(m) to gather, handle and analyse data and draw conclusions in ways that facilitate the attainment of the objectives of the enquiry;*
- *(n) to* ***demonstrate an understanding and appreciation of different cultures and universal values****; and*
- *(o) to* ***demonstrate empathy*** *in the handling of different issues.*

In alignment with the curriculum, the public examination design espouses the enquiry approach, which does not aim at assessing candidates' ability to identify predetermined "correct answers" (p.130) and multiple-choice questions are excluded from the assessment framework (p.132):

> *"Because Liberal Studies is concerned with the discussion and evaluation of issues, multiple-choice questions, as a kind of objective test, will not be adopted."*
>
> *(CDC & HKEAA, 2014, p. 132)*

The assessment framework of the LS Examination comprises two parts: the written examination and the School-based Assessment which is in the form of an Independent Enquiry Study (IES). In view of a much higher weighting (80%) on the written examination, empirical data elicited from the written examination alone will be regarded as representative for making judgement on the

validity of the assessment in this thesis.

The written examination is composed of Paper 1: a compulsory paper of three data-response questions; and Paper 2: a paper with three optional extended response questions, from which, candidates have to choose one to answer.

With the adoption of standards-referenced reporting in the grading of the HKDSE examination, the performance of candidates attaining Levels 1 to 5 in the examination is stipulated in the Level Descriptors (Table 1.2), which establish the basis for the grading process. To perform well in the assessment, in other words, to attain Level 5, candidates are expected to organise and analyse information from a diverse range of sources, to broaden their horizons in contemporary issues at the local, national and international levels, to evaluate various viewpoints and synthesise their own opinions by soliciting and conceptualising evidence. Levels 5* and 5** were determined by fixed percentages i.e. the top 10% and the next 30% of Level 5 candidates are categorised as Levels 5* and 5** respectively.

Table 1.2: Level Descriptors of the HKDSE LS Examination (HKEAA, 2014)

Candidates at this level typically:

Level 5	• show comprehensive knowledge and understanding of the key ideas and concepts of the subject by applying relevant knowledge and concepts to a diverse range of complex issues in particular contexts • identify relevant information, organise and analyse information from a diverse range of sources • interpret and analyse coherently the interdependence among personal, local, national and global issues from different perspectives • evaluate various viewpoints and synthesise their own opinions and suggestions on the basis of logical arguments and sufficient examples • communicate ideas in a concise, logical and systematic way • solicit and conceptualise evidence and show respect for evidence, demonstrating open-mindedness and tolerance towards a wide range of views and values • show initiative and self-management skills and reflect comprehensively and systematically throughout the enquiry learning process
Level 4	• show broad knowledge and understanding of the key ideas and concepts of the subject by applying relevant knowledge and concepts to a range of complex issues in particular contexts • identify relevant information, organise and analyse information from a range of sources • interpret and analyse the interdependence among personal, local, national and global issues from different perspectives • elaborate on various viewpoints and synthesise their own opinions and suggestions on the basis of logical arguments and some examples • communicate ideas in a logical and systematic way • solicit evidence and show respect for evidence, demonstrating open-mindedness and tolerance towards different views and values • show self-management skills and reflect comprehensively throughout the enquiry learning process
Level 3	• show general knowledge and understanding of the key ideas and concepts of the subject by applying relevant knowledge and concepts to some complex issues in particular contexts • identify relevant information, organise and interpret information from some given sources • consider and interpret appropriately the interdependence among personal, local, national and global issues from different perspectives • elaborate on viewpoints and give their own opinions and suggestions supported by arguments and some examples • communicate ideas in an organised manner • identify and show respect for evidence, demonstrating open-mindedness and tolerance towards different views • work with minimal reliance on teachers' instructions and reflect extensively on the implementation of the enquiry learning process
Level 2	• show basic knowledge and understanding of the key ideas and concepts of the subject by applying relevant knowledge and concepts to simple issues in particular contexts • identify relevant information • consider and interpret briefly the interdependence among personal, local, national and global issues from some perspectives • describe viewpoints and give their own opinions and suggestions supported by a few examples • communicate simple ideas • identify evidence, demonstrate tolerance towards particular views • work with some reliance on teachers' instructions and reflect broadly on the implementation of the enquiry learning process
Level 1	• show elementary knowledge and understanding of the key ideas and concepts of the subject by applying relevant knowledge and concepts to some simple issues in particular contexts • identify and gather some basic and simple information • identify simple relationships among personal, local, national and global issues from a few perspectives • list viewpoints and give some opinions and suggestions • express simple ideas briefly • identify and describe related information from their own viewpoints • work with detailed instructions from teachers and reflect briefly on the implementation of the enquiry learning process

The new core status in the Senior Secondary School Curriculum and the uniqueness of LS, when compared with subjects in the post-16 education of the U.K. and the U.S., have stirred up in society much concern about the high-stake assessment. In Hong Kong, the standard and requirements of the LS examination have not only been contentious among key stakeholders: teachers, students and parents, but also in the Legislative Council and in the media. The examination questions and requirements have made headlines of news reports ever since the first year of examination in 2012. In February 2017, the Panel on Education of the Legislative Council (Legislative Council Secretariat, 2017) discussed the implementation of LS. The level of difficulty of questions in the examinations, the issues examined, the assessment framework and the grading were among the topics for discussions by the councillors.

> *"9. Mr LAU Kwok-fan noted that frequent inclusion of questions on political issues (in the LS Examinations)"*
>
> *(Legislative Council Secretariat, 2017, p. 5)*

In the meeting, a non-binding motion for "the removal of compulsory questions from the LS public examinations and the adoption of a pass/fail grading system" (Legislative Council Secretariat, 2017, p.11) were finally passed.

In view of the social concerns of the subject, an evaluation of the validity of the examination is essential. Therefore, empirical data have been drawn from the 2015 LS Examination (HKEA, 2015) for illustrating an assessment validation process devised for large-scale examinations.

1.2 Structure of the Dissertation

The dissertation is structured as follows: In Chapter 2, the literature on the evaluation of assessment validity and the related cognitive models will be examined to justify the theoretical framework for the assessment validation process that will be deployed in this dissertation.

The adoption of mixed methods, underpinned by the pragmatic approach, will be discussed in Chapter 3. Besides, the data collection which was mostly based on secondary data from a joint study with the HKAGE, comprising a live script study, nominal group discussions and a think-aloud study, as well as the quantitative and qualitative analysis will be explained.

In Chapters 4, 5 and 6, the findings from the evaluation of the content and substantive validity of the 2015 HKDSE LS Examination will be discussed. The content validity evaluation in Chapter 4, in response to Research Question (1), will shed light on the compliance of the examination requirements with the expected performance stipulated in the assessment documents. Chapters 5 and 6 will be devoted to scrutinising the substantive validity of the examination. According to Messick (1995), substantive validity refers to the validity of the assessment of cognitive processes. Therefore, whether the cognitive processes were appropriately differentiated and whether candidates were assessed in terms of the requirements of the examination will be examined with reference to cognitive models: Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the New Taxonomy (Marzano et al., 2008) and Kuhn's (2001, 2005) KPI model.

In Chapter 7, the implications of the evidence-based multi-faceted assessment validation process adopted in the study for test developers and for the methodology of assessment validity research will be discussed. In addition, factors contributing to and limitations in terms of the applicability of the validation process will also be analysed.

CHAPTER 2 LITERATURE REVIEW

This chapter will start off with a review on the literature on the theoretical contentions about assessment validity and cognitive models that can be deployed in the validation process. Also, to justify the scope, significance and methodology of my research, the existing studies on the validity of large-scale assessments will be reviewed. Subsequently, the literature of Liberal Studies (LS) will be discussed to justify the choice of this examination for illustrating an assessment validation process of large-scale examinations.

2.1 Approaches to Assessment Validation

Assessment theorists, including Messick (1995) define validation as gathering evidence for "inferences about score meaning or interpretation and about the implications for action" (p.5). *Standards for Educational and Psychological Testing* (usually referred to as *the Standards*) (AERA et al., 2014) endorsed that assessment validation should hinge on the interpretation and use of assessment results. Under this premise, assessment validation is a process in which evidence is gathered "to provide a sound scientific basis for the proposed score interpretations" (AERA et al., 2014, p.11). Kane (2006) also concurred that the appropriateness of the measurement of the intended skills or knowledge should be evaluated with reference to the use of assessment results (as cited in Crisp & Shaw, 2012). Taking the HKDSE LS Examination as an example, the "interpretation and use" of the examination results is delineated in the Level Descriptors, which stipulate the requirements for each Level of Performance.

The evaluation of assessment validity can place emphasis on various aspects of the assessment process and take on different approaches. Among these approaches, content, criterion and

construct validity[6] have been extensively discussed in the literature of measurement theories (Messick, 1995; Ebel &Frisbie, 1991; Bloom et al., 1971, Pellegrino et al., 2016; Kane, 2013 and 2015; Newton et al., 2014) and will be explained in the following.

Approaches for evaluating **content validity** investigate whether the assessment is appropriate in measuring the content domains stipulated in the curriculum or course specifications. Even though measurement theorists, such as Kane (2013), criticised this approach for being subjective, the present study will still consider the "content" dimension since it is a significant element in assessment. Pellegrino & Wilson (2015) demonstrated the significance of "content" in their definition of assessment: a measurement of "what students are actually being taught", which should "parallel the curriculum" (p.264). Adhering to this definition, an assessment could only be considered valid if it aligns with the curriculum, which can be studied via a content analysis. The report by Johnson & Hayward (2009) on the benchmarking of LS with AQA A-Level Citizenship Studies provided insight to the content evaluation of the assessment of LS. They concluded that the HKDSE LS is comparable with the AQA A-Level qualification in terms of the curriculum and assessment demands. However, their report was based on an analysis of the curriculum, assessment framework and Level Descriptors, rather than the requirements of an authentic examination paper.

An evaluation of **criterion validity** focuses on how well the assessment results measure the "true value" (Cureton, 1951, as cited in Kane, 2013, p.18) of a criterion (such as an attribute). APA et al. (1966) put forth that criterion-related validation is a comparison of "the test scores with one or

[6] *Various terminologies have been adopted by measurement theorists. Pellegrino et al. (2016) summarised that cognitive validity is a variety of Messick's construct validity and "traditional content and consequential validity, and inferential validity is related to criterion validity" (p. 62). Some theorists adopted a finer approach in the classification. For instance, the first edition of the Standards (1985, as cited in Pellegrino et al. (2016)) distinguished between predictive and concurrent validity with regard to the time for making inferences.*

more external variables considered to provide a direct measure of the characteristic or behaviour in question" (as cited in Newton et al., 2014, p.106). Statistical methods are employed in determining criterion validity. For instance, a correlation of the assessment scores and the external criterion measures or an expected value of the attribute can be conducted in the validation process (Pellegrino, 2016, p.61). Nevertheless, this approach of validity evaluation has been contested by a number of measurement theorists, such as Ebel (1965), who questioned the feasibility of identifying the "true value" of the criterion. Kane (2013) also quoted the comments of Cronbach (1971) and Guion (1998) on the difficulty in acquiring "a good criterion" (p.19).

In view of the criticisms towards the criterion-based approach, Cronbach & Meehl (1955) propounded **construct validity**. In their terms, assessments should be evaluated with respect to the intended construct of the assessment (such as, the thinking skills) (as cited in Newton et al., 2014) by deploying cognitive models. Along a similar vein, Pellegrino et al. (2016) also suggested that assessment validation should be based on "evidence of students' competencies" (p.63). Their argument was premised on their earlier piece of work on the Assessment Triangle (Pellegrino et al. 2001), which put forth that the cognitive skills of candidates can be observed in assessments as "evidence" for further interpretation.

Other than these three dominating approaches to assessment validation, Newton and Shaw (2014) postulated the social aspect as another dimension for assessment evaluation. According to them, public acceptance and the fairness of an assessment should also be validated. Nevertheless, the evaluation of the social consequences of assessments has aroused much contention among measurement theorists. Examples of the opposing views are put forth by Maguire, Hattie and Haig (1994); Mehrens (1997); and Popham (1997) (as cited in Shaw et al., 2012, p.171). I concur with Shaw et al. (2012) that the consequential dimension is "external" (p.171). Although Hong Kong

is akin to the Asian countries in the study of Manns (2018) with a prevailing "culture of testing" (p.13), which attaches much importance to public examinations, evidence on the substantive aspects, rather than that on the social impacts of the examination, may shed light on the validity. Evidence justifying the validity in the substantive aspect (i.e. the theoretical (cognitive) processes adopted in the responses) may also address the social concerns over the appropriateness of the assessment. Therefore, this study will not directly probe the "external" social or consequential dimension.

Instead of delving into the contention of the focus of assessment validation, Messick (1995) unified various approaches under construct validity. He advocated the evaluation of construct validity by integrating evidence from various aspects, including content, substantive, generalisability, external and consequential components. In his words, validation entails "ascertaining the degree to which multiple lines of evidence" support the inferences (as cited in Moss et al., 2006, p.115). Adopting Messick's perspective, multiple sources of evidence on the content and substantive aspects were elicited to provide a stronger basis for evaluating assessment validity in this study.

Furthermore, the current study aims to devise a validation procedure that might be deployed for any large-scale examination, not being specific to certain examination boards. The LS Examination will be used purely for illustrating the procedures. The availability of evidence will be a determining factor in selecting the aspects of validity to be incorporated. Since the scores of all candidates in the 2015 LS examination and an external measurement equivalent to LS are not available, the generalisability of the validation results and the external aspect of validity were not part of the research objectives. Instead of gathering evidence specifically for the consequential aspect, this research focuses on the content and substantive aspects.

Apart from Messick (1995), Goldstein (2015) also suggested academics should move away from the contention of the nature of validity and "provide empirical evidence for the things" "associated with" and "undermining" validity (p.198). Exploring further into the methodological aspects, Kane (2013, 2015) postulated the Argument-based Approach for gathering evidence into the evaluation of construct validity, which was also recommended by AERA et. al (2014):

> *"Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use." (p.11)*

The Argument-based Approach is an evaluation under an interpretive argument and validity argument. The former delineates the "interpretation and use" of the assessment and might comprise "four inferences" (Kane, 2013, p.22) (Table 2.1), though the number of inferences may vary with the nature of the assessment:

Table 2.1: Inferences of interpretive arguments suggested by Kane (2013, p.22)

1. *"Scoring"* 2. *"Generalisation: …an estimate of the expected score over a universe of similar performances…"* 3. *"Extrapolation: …inference …to expected performance in a domain of non-test performances."* 4. *"Decision making: …a decision about the test taker (e.g. certification)"*

The interpretive argument pins down the type of evidence for the justification of the inferences, as specified by the validity argument. Nevertheless, the necessity of the interpretive argument, especially for a public assessment, has been subjected to a heated debate among theorists, such as Newton and Serici (as cited in Kane, 2013), because "the interpretation and use" of the assessment is stipulated in the curriculum or syllabus. In response to Newton and Serici's call for focusing on the validity argument alone, Kane (2013) conceded that the interpretive argument may not be necessary provided that the "interpretation and use" of the scores is clearly stated. For the case of the LS Examination, the Level Descriptors of the Diploma, developed based on the Assessment

Objectives and the curriculum aims, stipulate clearly the performance of candidates achieving various levels. Therefore, it is not fundamental to delineate a set of interpretive arguments. The justification of the validity argument should be based on the evidence gathered with reference to the Level Descriptors.

For the evaluation of the validity argument, theorists such as Gorin (2007) suggested studying the internal theories to the construct (i.e. "the cognitive and metacognitive strategies, and multiple alternative paths") (as cited in Newton et al., 2014, p.150). Cognitive models can be deployed to examine the knowledge and skills used to justify the validity argument. Similarly, Newton et al. (2014) also advocated the validity study of assessments in relation to the construct. Pellegrino et al. (2001) believed that assessments provide evidence for the cognitive skills of candidates[7]. In Kane's (2013) terms, assessments demonstrate "whether they use what they know" (p.62). Sireci (2007) also concurred with the deployment of evidence in assessment validation (as cited in Crisp and Shaw, 2012, p.221):

> *"If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose."*

All the above literature converged to the justification of the validity argument by being based on evidence of the alignment of the actual performance of candidates in an examination with cognitive models.

In light of the validity approaches in the literatures, the Argument-based Approach (Kane, 2013,

[7] *Pellegrino et al. (2001) proposed the Assessment Triangle as follows:*
"a model of student 'cognition and learning' in the domain, a set of beliefs about the kinds of 'observations' that will provide evidence of students' competencies, and an 'interpretation process' for making sense of the evidence." (p.44)

2015) constitutes the evaluation framework for the content and substantive aspects of construct validity in my study. The validity of the 2015 HKDSE LS examination was evaluated, with reference to the knowledge and skills assessed, by employing cognitive models. The differentiation of the cognitive ability of candidates by the stipulated Levels of Performance was also examined by analysing authentic scripts.

2.2 Empirical Studies on Assessment Validity

As explained in Section 2.1, there are insufficient studies on the assessment of LS in the literature. Even though public examination subjects with Assessment Objectives similar to LS are found in the U.K., only a few research studies have been carried out on the validity of these examinations. In this section, the literature on the validity of large-scale assessments of other subjects adopting the Argument-based Approach will be discussed to shed light upon the methodology of my study.

In the literature of assessments in general, the locus of research has shifted from the procedural aspects of test developments to test validity since the late 1980s. However, according to a recent study by DeLuca (2011), the emphasis of scholars has been on the theoretical and epistemic aspects of validity mainly, leaving a gap in the literature on the link between the "contemporary theoretical foundations in validity to practical methods" (p. 304) of assessments. Moss et al. (2006) and Shaw et al. (2012) also observed a gap of validity literature in the operational framework. DeLuca (2011) proposed integrating the "dialectic, hermeneutic and transgressive forms of enquiry" (p.303) in the argument-based structure of validity studies. His work illuminated the methodological approach of assessment validity studies, which awaits empirical studies to operationalise the methodology.

In the scarce literature on the methodology of assessment validity, studies by Mahon (2006), Mason (2007), Zahedi et al. (2012), Chapelle (2012), Crisp & Shaw (2012), Shaw & Imam (2013) and Manns (2018) are among the few examples of validity research of large-scale high-stake assessments. The research of Manns (2018), though on high-stake assessments in Asia, focused on the socio-cultural impacts rather than substantive validity. All of the studies above except the study of Shaw & Imam (2013) were based on quantitative analysis of scores, not fulfilling the requirement for validation of assessments from multiple sources of evidence as put forward by Messick (1995) and Cook et al. (2016). Shaw & Imam (2013) exemplified a validation study on qualitative data by a text analysis of scripts by examiners, offering an alternative methodological approach on construct relevance. However, the focus of their study was on the use of English in some content-based examinations, rather than the substantive validity, which is the focus of my research.

Chapelle (2012) supported Kane's Argument-based Approach for the evaluation of the validity of language assessments. She advocated the use of the interpretive and validity arguments to formulate a clear definition of problems, leading to a more focused validation study. Crisp & Shaw (2012), also adopting the Argument-based Approach, conducted a quantitative Rasch analysis on the generalisation from the examination scores of international and GCE A-Level Examinations.

Crisp & Shaw (2012) deemed qualitative evidence "problematic" (p.200) for the validation of examinations. Newton et al. (2014), quoting Shaw & Crisp (2012), commented that the challenge in gathering qualitative evidence is that it is time-consuming to collect and the validation process might not be a one-off study if qualitative in nature.

However, being time-consuming in terms of data collection is not a justifiable deterrent for

research interests with qualitative data. Crisp & Shaw (2012) conducted a study on the construct validity of international A Level Geography examinations using a mixed methodology approach: Rasch analysis and qualitative analyses on the expected and authentic processes adopted by candidates and the potential problematic items as identified by the Rasch analysis. Their research illustrated that both quantitative and qualitative evidence could be elicited for assessment validation. Lim (2014) was also dubious about the emphasis on psychometric research in the territory of validity evaluation of large-scale assessments. In lieu of examination scores, she conducted a study on the cognitive processes that candidates underwent when taking the TOEFL iBT reading test. Though her study was using quantitative scores on the cognitive processes as evidence, it did not rely upon psychometric statistics of the scores.

Taking an approach similar to these research studies, my current study evaluated quantitatively and qualitatively the validity of the 2015 LS Examination with reference to the authentic skill performance of candidates, instead of focusing on the examination scores, to eliminate the factor of marking quality and to provide direct evidence for the various Levels of Performance as stipulated in the Level Descriptors.

Gathering evidence for validating large-scale examinations from performance data of candidates and the views of examiners/ experts on the cognitive demands of the examination was also advocated by Shaw et al. (2012). In their study, even though the data were quantitative in nature (the ratings by examiners), they also believed that the answer scripts of candidates could be analysed qualitatively to study the validity argument of whether “the tasks elicit performances that reflect the intended constructs” (Shaw et al., 2012, p.168). Linn et al. (2000) also shared the view that cognitive processes could be evaluated in research on construct validity. Weinstock (1999) conducted a qualitative analysis on the cognitive processes of juror reasoning by deploying

the scoring criteria developed by Kuhn et al. (1994) (as cited in Kuhn, 2001). The thematic analysis in terms of "the representation of verdict criteria", "the use of evidence" and "the relation of evidence to verdict" (p.2) provided insight for the qualitative study of the argument formulation skills of candidates as shown in the answer scripts in the present dissertation.

Besides cognitive skills, such as reasoning, evidence of the application of knowledge and metacognitive skills can also be solicited in assessment validation. In Kuhn's (2001) terms, "there is much more than needs to develop the procedural skills themselves that enable people to acquire new knowledge" (p.7). Meta-level management and values also determine knowledge acquisition, which should be deployed in assessments. Even though the research studies by Kuhn (2001) and Weinstock (1999) (as cited in Kuhn 2001) focused on knowledge acquisition processes rather than on the validity of the assessment of cognitive skills, their categorisation of various levels of reasoning/ judgement-making skills, the meta-level and values/ knowledge application procedures purported could be applied in the evaluation of the performance of candidates in examinations.

In a nutshell, to fill the research gap, a validation process for large-scale exainations focusing on the actual performance of candidates and providing direct evidence for the performance differentiation and the interplay among knowledge, values, argument-formulation and metacognitive mechanisms will be put forward in the present study.

2.3 Research on LS

Based on a search of the research interests in the field of "Liberal Studies", "assessment" or "examination" in "Hong Kong, via 'Google Scholar', the University of Hong Kong Libraries and University of Bristol Library[8], this dissertation takes a different direction from the existing research studies on LS, which are predominantly inclined to the teaching and learning of the subject.

The studies on LS focused mainly on the curriculum and pedagogy on integrating knowledge of various disciplines and critical thinking skills. Before the curriculum of LS came into place in 2009, a number of studies ((Morris and Chan, 1997, A.W. L. Leung, 2009, Deng, 2009) focused on the integrative, cross-disciplinary nature of the curriculum of LS. Studies on the pedagogy of LS included Cheung's (2009) research on integrating media education in LS lessons and the research on LS teacher training by Lai and Lam (2011). Fung and Howe (2012) studied the effectiveness of collaborative group work in teaching critical thinking in LS. Fung (2016) conducted a five-year longitudinal study via a documentary enquiry of 560 newspaper articles on the perceptions of students and teachers on LS before and after studying the subject. He concluded that students demonstrated "favourable attitudes towards LS before studying the subject". However, after the implementation of the LS curriculum, students expressed "a certain degree of disappointment" and teachers lacked confidence in teaching the subject, criticising the curriculum as "overly ambitious" (p.625) and too broad.

A recent addition to the studies of the learning outcomes of LS was conducted by Chiu et al. (2018). They studied the impact on LS of students' civic values and engagement in Hong Kong

[8] *With the key-word search of "Liberal Studies", "assessment" or "examination" in "Hong Kong" via Google Scholar, the University of Hong Kong Libraries and University of Bristol Library, 2690, 372 and 21 results were generated respectively.*

society, placing emphasis on the learning impact rather than the assessment of LS.

Even though these studies did not have direct relationship with the assessment of LS, they were pointing to the nature of LS as being multi-disciplinary in knowledge basis and a subject for the development of critical thinking, which are the key assessment elements to be validated in my study. In view of the fact that the validity of public examinations in Hong Kong has been receiving relatively less attention among researchers, an empirical study of the authentic performance of candidates (including live examination scripts and a think-aloud study) has been conducted in this dissertation to shed light on the validity of the 2015 LS Examination.

Curricula equivalent to the HKDSE "Liberal Studies" could not be found in the pre-university levels in the UK or US. Therefore, studies of the assessments of similar subjects are rare. The curricula that shared the most similar curriculum aims and Assessment Objectives with LS were AQA A-Level General Studies and AQA A-Level Citizenship in the UK[9]. All these curricula aim to broaden students' knowledge and understanding of contemporary issues; and develop thinking skills as shown in the quotations below:

> ***HKDSE LS:***
>
> *"To enable students to develop multiple perspectives on perennial and contemporary issues in different contexts' and 'to develop students in a range of skills for life-long learning, including critical thinking skills."*
>
> *(The Curriculum Development Council (CDC) and HKEAA, 2014)*

[9] *The candidature of these subjects in the UK is far lower than that in the HKDSE. In 2018, the candidature of HKDSE LS was 53 691, AQA AL General Studies 2 313 and AQA A-Level Citizenship Studies 287. Pearson Edexcel ceased to offer A-Level General Studies in 2011.*

AQA A-Level General Studies:

"To encourage thinking across specialist subjects, the cultural, scientific and social domains are divided into 'Culture and Society' in Modules 1 and 3 and 'Science and Society' in Modules 2 and 4."

(AQA, 2014, p.2)

Pearson Edexcel A-Level General Studies:

"To integrate knowledge from a range of disciplines in order to develop an understanding of the interrelationship between them and encourage a broader and deeper understanding of issues" and "to think logically and creatively in order to assess the relative merits of evidence"

(Pearson Edexcel, 2014, p.10)

AQA A-Level Citizenship Studies:

"To develop a critical interest in topical citizenship issues and debates, and encourage independent thinking skills"

(AQA, 2014a)

Despite the similarities in the curriculum and assessment aims, the assessment frameworks are different. The written examinations of AQA A-Level General Studies comprise multiple-choice questions. Resembling the examination of the HKDSE LS, the AQA A-Level Citizenship Studies examination consists of data-response questions and essay-writing. Citizenship Studies was offered in some European countries. In the literature, there were only a few research studies on the assessment of Citizenship Studies, such as the research conducted by Kerr et al. (2009). Their study compared "the nature and effectiveness of approaches" (p.86) to the assessments in citizenship education for ages between 5 to 18 in European countries from 2005 to 2008. The research was based on policy studies and documents, meetings soliciting views from experts from the participating European countries. Empirical performance data on the effectiveness of the assessment of the subject were not deployed.

In reviewing the literature on courses at the university level, a number of courses which bear names akin to Liberal Studies were found, for instance, Liberal Education in Miami University in the 1980s (Schilling, 1987) and Liberal Arts or Interdisciplinary Education at K-12 and university levels in the U.S. Among these subjects, Liberal Arts Studies are more widely different from the HKDSE LS as they are discipline-based. Some Liberal Education programmes are interdisciplinary. In the comparison research of Rhoten et al. (2000) on the assessments of interdisciplinary and discipline-based courses, they put forth an argument that performance-based assessments are "suited to measure the complexity, ambiguity, and multiplicity of skills" (p. 4), which are the Assessment Objectives of interdisciplinary courses. In their view, the assessment of an interdisciplinary course should aim at assessing students' ability "to integrate" "multiple disciplines" (p.18) and "to put such capacity to use" (p.16) in performance-based tasks. Although their study focuses on courses at the university level, rather than the secondary school level, it provides insight into the assessment formats appropriate for measuring the performance of students of integrated interdisciplinary courses, which are of a similar nature to LS. Their findings also lend support to the use of performance-based authentic assessment in the LS examination. However, their suggestions on the appropriateness of assessment methods were based on literature review, questionnaire surveys and interviews with course providers. Contrary to their approach, the collection of empirical evidence for assessment validation was advocated by Messick (1995) and Goldstein (2015). Following this line of thinking, my study was an empirical study on the "actual consequences of an assessment" suggested by Messick (1989) (as cited in Shaw et al., p.162).

Since the examination of LS has been conducted for only eight years in HK, there have been few empirical studies specifically on the examination of Liberal Studies. One of them was conducted by the HKEAA (2007). This study aimed at developing the Level Descriptors of the standards of

the HKDSE LS Examination through the analysis of the actual performance of candidates in the 2005 Advanced- Supplementary (AS) Level LS Examination. At the time of the study, candidates taking the HKDSE LS were not available. Therefore, it is worthwhile and practicable to conduct an empirical study of the actual performance of HKDSE candidates to investigate the appropriateness of the Level Descriptors in the differentiation of the performance of candidates. Furthermore, the study in 2007 clearly delineated the component parts of the Level Descriptors (namely, "Knowledge and Understanding", "Generic Skills" and "Enquiry Competence" (HKEAA, 2007, p.3)), providing a framework of skill domains for analysing candidates' performance in this current study.

Apart from this official empirical study, there are a few academic articles focusing on the marking of LS examination. Examples are the studies conducted by Coniam (2011), Coniam & Yeung (2010) and Coniam & Falvey (2016) on the impact of On-screen Marking and double-marking[10] on the marking standard of Liberal Studies by comparing the marks awarded by examiners via the on-screen marking system and that on paper. Kuo (2007) took a different approach from an empirical study. By a theoretical logical comparison between analytic and holistic marking rubrics, he concluded that analytic marking rubrics should be adopted instead of the holistic one, which is stipulated in the Assessment Framework.

Undeniably, the marking process is one of the factors affecting the validity of an examination. However, the marking guidelines and the marking standard of markers will not be the focus of this study as they are question-specific and vary from year to year. In this study, the permanent and ultimate requirements of the diploma as stipulated in the Level Descriptors will be evaluated

[10] *Markers of LS mark scripts at the marking centres via the workstations. Every script is marked by at least two markers. If there is a large discrepancy of marks awarded to a script by two markers, the marking system will allocate the script to another marker for marking. The highest and closest pair of marks will be the final mark of the script.*

by analysing the actual performance of skills and knowledge application of candidates in the LS HKDSE examination. To minimise the variation in the scoring standards and quality of markers, experienced examiners were invited to take part in the re-scoring process of the live script analysis.

Research studies have also been found on the perception of stakeholders on the assessment of LS. In 2018, the Hong Kong Federation of Youth Groups conducted a study on the views of students and teachers on the LS assessment and found that 59% of teachers and only 28.4% of students believed the LS examination reflected the ability of candidates effectively (HKFYG, 2018, p.7, 12). However, this study was an opinion survey rather than academic research. Hitherto, research most relevant to the field of the current study was conducted by L. S. Leung (2013, 2017)[11]. In 2013, Leung studied teachers' perception on the relationship between teaching strategies on higher-order thinking skills and the LS public examination by conducting a questionnaire survey of 41 teachers followed by 12 in-depth interviews among the survey respondents. Teachers were found to adopt examination-oriented strategies in teaching higher-order thinking skills in LS. Adopting a similar methodology, in 2017 L. S. Leung conducted a study on the alignment of the LS public examination and the curriculum aims. Based on a questionnaire survey of 42 schools, 35 semi-structured interviews with students, teachers and policy-makers and 10 classroom observations, L. S. Leung concluded that the LS public examination did not fully align with the curriculum aims for nurturing 21st Century skills. Along a similar vein of her own study in 2013, L. S. Leung (2017) contended that the domination of the exam-oriented teaching and learning strategies in LS led to a failure to nurture comprehensively 21st Century skills:

[11] *Information of L. S. Leung's research in 2017 was based on her Powerpoint Presentation in a seminar conducted on 11 March 2017 in Hong Kong.*

- *(The examination does) "not really assess the outcome from students' authentic processes of learning but mainly the abilities of applying procedural knowledge of argumentative issues";*
- *"the consistent patterns, mental models and question forms that have been asked enable students to bypass critical thinking, produce seemingly sensible judgement on issues asked (in the examinations) without considering fundamental reasoning under question";*
- *"only focus (on) learn(ing) to think restrictively on individual but not on independent critical thinking". (L. S. Leung, 2017, Slide 25)*

The focus of L. S. Leung's study in 2013 was different from my study. Here interest is on the validity of a public examination instead of the backwash effects of the public examination on teaching. The focus of L. S. Leung's research in 2017 was closer to mine. The alignment of the public assessment with the curriculum aims is one of the aspects in a validity study of an assessment as suggested by Pellegrino & Wilson (2015). (The value of an evaluation of content validity has been discussed in Section 2.1.) However, L. S. Leung (2017) studied the alignment by investigating the impact of the public examination on the teaching and learning strategies. Taking on a different perspective, my research is centred around whether the 2015 LS public examination appropriately differentiated and reflected the thinking skills and knowledge of candidates as stipulated in the Assessment Objectives. To fulfil this research purpose, the perceptions of teachers and students on the impact of assessment on "learning to think" (Table 2.2) cannot provide hard evidence for the actual thinking skills and knowledge of candidates. Therefore, in my study, live examination scripts, think-aloud protocols and views of examiners on the authentic performance of candidates formed the basis of evidence for justifying the validity of the HKDSE LS Examination in assessing the thinking skills of candidates. The criticisms on the inadequacies in the thinking skills of LS students, as well as a comparison between the perceptions of teachers and students on the thinking skills stipulated in the LS curriculum discussed by L. S. Leung (2017) and the actual performance of candidates in the examinations have also been examined in this research.

Table 2.2: Questions in the survey on the outcomes of "learn(ing) to think" in L. S. Leung's research (2017, Slide 15)

> *"Please indicate the extent to which you agree that each of the following statements is the aim of the Liberal Studies (LS) curriculum and the extent to which the DSE helps to achieve these aims.(On a 1 to 4 scale, with 1=most negative and 4= most positive)*
>
> *e. Nurture rational, objective and critical thinking skills*
> *f. Nurture innovation and the ability to solve problems effectively*
> *g. Help students become independent thinkers through developing their skills in knowledge construction and issue-enquiry"*

Literature related to the cognitive models in the theoretical framework of the validity evaluation process adopted in this study will be discussed in the following section.

2.4 Cognitive Models for Assessment Validation

As discussed in Section 2.1, Kane (2013, 2015) suggested the evaluation of construct validity by investigating the alignment between the performance of candidates and cognitive models. In this section, the cognitive models to be deployed in the proposed validity evaluation process for the 2015 LS Examination will be discussed.

2.4.1 Bloom's Taxonomy

As evidenced by Google search, Bloom's Taxonomy remains the most influential cognitive model and the most frequently cited in academic writings on educational measurements ever since its development in the late 1950s. Therefore, Bloom's Taxonomy is chosen for evaluating the appropriateness in the differentiation of performance on thinking skills in the LS examination in this study.

In *Handbook I,* Bloom (1956) established a common framework for test development by

classifying the educational goals or outcomes related to cognitive skills into a hierarchy. Six cognitive skills were identified and organised into a hierarchy of intellectual demands (Bloom, 1956) (Table 2.3):

Table 2.3: Bloom's Taxonomy

Level 1. Knowledge
Level 2. Comprehension
Level 3. Application
Level 4. Analysis
Level 5. Synthesis
Level 6. Evaluation

Both the intellectual demand and the complexity of the cognitive skills required are increasing in ascending order from Level 1 to Level 6. The command of the lower level objectives facilitates that of the successive levels, though it may not be a necessary precursor.

In Handbook II, the affective domain[12], including interests, appreciations, attitudes, values, and adjustments to them, was added to the educational objectives. Anderson et al(1994) suggested that educational objectives regarding students' interests, values and attitudes could be developed. However, the fulfillment of the objectives in the affective domain is often dependent on the achievement of the objectives in the cognitive domain. Under the assumption that the values and attitudes of candidates are involved in the cognitive processes, such as evaluating and synthesising, when they work on the assessment tasks, the cognitive domain should therefore be the basic framework for analysing the performance of candidates in the present study.

Many schemes for organising cognitive skills have been developed since Bloom's Taxonomy. However, the significance of Bloom's Taxonomy is indisputable among theorists. For instance,

[12] *The hierarchical organisation of the different levels of the education objectives in the Affective Domain as suggested by Krathwohl et al. is: 1. Receiving (Attending), 2. Responding, 3. Valuing, 4. Organisation, 5. Characterisation (as cited in Anderson et al., 1994).*

Leighton (2011) commented that "an internet search of *higher-order thinking"* "ties the origin of the term to" Bloom's Taxonomy (p.155). Marzano et al. (2008) acclaimed the taxonomy as a model that makes "a major contribution to the science of designing educational objectives" (p.1). Leighton and Gierl (2011) also recognised the taxonomy as a guide to the development of educational test items for assessing various levels of cognitive complexities. Being the origin of cognitive models, Bloom's Taxonomy is an incontestable choice as the basis for evaluating the differentiation of cognitive skills by the examination.

The hierarchical organisation in Bloom's Taxonomy is supported by empirical studies, making it a credible basis for evaluating the assessment of cognitive skills. The meta-analysis[13] conducted by Kreitzer and Madaus (1994) was an example of these studies (as cited in Anderson & Krathwohl, 2001)). They found out that students scoring higher marks in tests on *Comprehension* did not attain equally high scores in tests on a higher-level skill--*Analysis*. Their study offered evidence for the hierarchical ordering from *Comprehension* to *Analysis* with increasing cognitive complexity. Further evidence to the hierarchical ordering of *Comprehension*, *Application* and *Analysis* was elicited from the studies of Hancock (1998) on Multiple Choice Tests and Constructed-response Tests (as cited in Anderson & Krathwohl, 2001).

Despite being widely cited in research on the development of educational objectives and assessment designs, Bloom's Taxonomy also drew criticism and tremendous interest among multitudinous academics leading to its revisions and the development of new taxonomies, employing different terminologies, domains, ordering and dimensions.

[13] *In these studies, students took tests on the six categories of skills in the taxonomy. Correlations among the scores in the six levels of the taxonomy were analysed in a two-way matrix.*

Various academics, including Anderson & Krathwohl (2001)[14], and Biggs (1982) identified some issues with Bloom's Taxonomy. Leighton & Gierl (2011, p.54) pointed out the key critiques in relation to the definition of cognitive skills and the lack of consideration of individual variations and empirical evidence:

> *"…the cognitive skills are not defined explicitly…"*
> *"…there is little or no cognitive-response variability in how students solve the items…"*
> *"…there is no empirical evidence, i.e. the test score inferences about complex cognition are weak."*

The debates over the hierarchical ordering of cognitive skills persist due to the lack of empirical studies on students' performance to provide evidence for the ordering. The ordering of *Evaluation* and *Synthesis* in the hierarchy is one of the most contentious themes. According to the meta-analysis by Kreitzer and Madaus (1994) (as cited in Anderson & Krathwohl, 2001), although *Analysis* was found to be further lower than the demand of *Synthesis* and *Evaluation*, the differences in the correlation between the scores for *Analysis* and *Evaluation* and that for *Analysis* and *Synthesis* were insignificant and conclusion could not be drawn on the discrepancies in levels of cognitive complexity between *Evaluation* and *Synthesis*. Incongruent to Bloom's Taxonomy, Hauenstein (1998) supported the reversed ordering of *Evaluation* and *Synthesis*[15].

The status of *Knowledge* in the thinking process is also the epicentre of controversies. The findings by Kreitzer and Madaus (1994) lent support to the possibility for *Knowledge*, including factual,

---

[14] *The drawbacks of Bloom's Taxonomy pointed out by Anderson & Krathwohl (2001) are summarised as follows:*

- *"There are different interpretations of cognitive skills. For instance, some academics, like Orlandi (1971) argued that comprehension involves analysis."*
- *"The linear relationship among the categories may not exist. The lower levels of skill may not be a pre-requisite for the higher levels. Besides, the cognitive complexity of the six categories may not be ordered as in Bloom's Taxonomy. The most controversial aspect of the ordering is whether the cognitive complexity of synthesis is lower than that of evaluation." (T. Y. G. Leung, 2017)*

[15] *Hauenstein (1998) found that Analysis can be a subcategory of Evaluation, but Synthesis is independent of Evaluation.*

procedural and conceptual aspects, to be categorised separately from cognitive skills (as cited in Anderson & Krathwohl, 2001). Anderson et al. (1994) commented that the acquisition of *Knowledge* involves not only the mastery of concepts, but also skills like reasoning. They concurred that certain demands for *Knowledge* may be more complex than those for higher levels in the Taxonomy.

In view of these controversies in the hierarchical ordering of skills, Bloom's Taxonomy was not the only cognitive model to be deployed in the assessment validation process in this thesis. The choice of the Revised Taxonomy by Anderson & Krathwohl (2001), the New Taxonomy by Marzano et al. (2008) and Kuhn's (2001, 2005) KPI Model will be explained in the following parts.

2.4.2 Taxonomies after Bloom's

After Bloom's work, various schemes of organising intellectual skills expected of students were formulated. As classified by Anderson & Krathwohl (2001), these schemes could be unidimensional[16] and multidimensional systems.

The SOLO taxonomy by Biggs (1982) is an example of a unidimensional system in which conceptual structures are classified into four levels according to their complexity: 1. Unistructural, 2. Multi-structural, 3. Relational and 4. Extended Abstract. However, it is more applicable to the marking of the "structural complexity" (Biggs, 1982, p.178) of responses to assessments. For the evaluation of the differentiation of cognitive skills (not only the "structural complexity") by

[16] *Unidimensional models rank cognitive skills under a single hierarchy. They can be exemplified by Gagne (1972), Gagne & Briggs (1979), Biggs & Collis (1982), Hauenstein (1998)(as cited in Anderson & Krathwohl, 2001, and Webb, 1997, 2007).*

assessments, Bloom's Taxonomy and its families are more practicable as the relative cognitive demands of thinking skills to be assessed are specified.

Depth of Knowledge (DOK) suggested by Webb (2002) is another more recent addition to the family of unidimensional models. Webb (2002) stated that DOK is a taxonomy for checking against "if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards" (as cited in Lane, 2005, p. 7). Wyse and Viger (2011) termed it as an alternative to Bloom's Taxonomy. The DOK describes "the cognitive demands and complexity required by an item" (Wyse and Viger, 2011, p.188) in four levels: 1. Recall; 2. Skill/ Concept; 3. Strategic Thinking and 4. Extended Thinking. Webb (2002) categorised "evaluating solutions to problems" and "using concepts to solve problems" under Level 3 Strategic Thinking; whereas "making predictions with evidence as support" and "synthesising information from multiple sources" as the most cognitively demanding, Level 4 Extended Thinking, reversing the hierarchical order of Bloom's Taxonomy. In comparison with Bloom's Taxonomy (1956), the DOK places emphasis on the application of subject-specific knowledge and concepts in answering the questions. As a non-content-based subject, the mastery of cognitive skills, rather than the use of a set of subject-specific concepts, is the key assessment criteria in LS. Therefore, Bloom's Taxonomy, which focuses on the complexity of cognitive skills, irrespective of the specific knowledge or concepts applied, will likely be more appropriate as a tool for evaluating the validity of the LS examination.

In the present study, the most frequently cited and recent cognitive models were chosen as the conceptual framework for analysis. As such, the Revised Taxonomy (Anderson & Krathwohl, 2001), which is closest to the origin of all taxonomies, Bloom's Taxonomy (1956), and a more recently developed multidimensional system, the new taxonomy by Marzano et al. (2008), were

selected as the cognitive models for the validation of the 2015 LS Examination.

Anderson & Krathwohl (2001) revised Bloom's Taxonomy[17] in light of critiques. In their revision, *Knowledge* is separated from cognitive skills as a distinct dimension. Even though different terminology was adopted, the hierarchy of cognitive skills remained similar. The most prominent difference between Bloom's Taxonomy and the revised one lay in the rankings of *Synthesis* and *Evaluation* (which are termed *Create* and *Evaluate* respectively in the Revised Taxonomy) (Table 2.4). The rankings of these two skills were swapped in the two taxonomies.

Table 2.4: The Revised Taxonomy (Anderson & Krathwohl, 2001)

Level 1. Remember
Level 2. Understand
Level 3. Apply
Level 4. Analyse
Level 5. Evaluate
Level 6. Create

The New Taxonomy of Marzano et al. (2008) is a relatively new development of cognitive models. It is two-dimensional, with separate hierarchies for *Processing* and *Knowledge* (Tables 2.5 and 2.6). Being different from Bloom's Taxonomy, while similar to Anderson & Krathwohl's (2001) revised version, the New Taxonomy sorts out *Knowledge* from the mental processes as a distinct domain.

[17] *Anderson & Krathwohl (2001) did not classify their system as multi-dimensional and the classification was not presented in a multi-dimensional manner. However, I group Anderson & Krathwohl's classification system under multi-dimensional as there is a separate dimension for Knowledge. According to Anderson & Krathwohl (2001), Knowledge goes from Remember, Factual Knowledge, Conceptual Knowledge to Procedural Knowledge in ascending order of complexity. Schraw & Robinson (2011) also described this classification system as "two dimensional", which "brings it more in line with the psychological research literature on the underlying structure of thinking" (p.158).*

Table 2.5: A comparison between the New Taxonomy and Bloom's Taxonomy[18]

Levels of Mental Processing *(Marzano et al., 2008)*	***Similar Levels in Bloom's Taxonomy***
Level 1. Retrieval	*Level 1. Knowledge*
Level 2. Comprehension	*Level 2. Comprehension*
Level 3. Analysis	*Levels 4, 5, 6. Analysis, Synthesis, Evaluation*
Level 4. Knowledge Utilisation	*Level 5. Synthesis*
Level 5. Metacognitive System	*Nil*
Level 6. Self-System	*Nil*

Table 2.6: Domains of Knowledge (Marzano et al., 2008)

Knowledge domain	***Hierarchical component***
Information	1. *Vocabulary terms* 2. *Facts* 3. *Time sequences* 4. *Generalisations* 5. *Principles*
Mental Procedures	1. *Single rule* 2. *Algorithms* 3. *Tactics* 4. *Macroprocedures*
Psychomotor Procedures	1. *Foundational procedures* 2. *Simple combination procedures* 3. *Complex combination procedures*

Incongruent to Bloom's taxonomy, *Generalising*, *Specifying* and *Evaluating* (which are Levels 4, 5 and 6 as classified by Bloom) were all grouped under *Analysis* (Level 3) in the New Taxonomy. Furthermore, some new mental processing levels were specified in the hierarchy: *Knowledge Utilisation* (using knowledge to accomplish a specific task, especially authentic tasks) replacing *Synthesis* (Level 5 in Bloom's); *Self-system* (involving attitudes, beliefs, behaviours that control motivation) and *Metacognitive* Systems (focusing on setting and monitoring goals, including *Specifying Goals*, *Process Monitoring*, *Monitoring Clarity* and *Monitoring Accuracy*) as levels

[18] *According to Marzano et al. (2008), Level 3 Analysis matches with Level 4, 5 and 6 in Bloom's taxonomy, whereas Level 4 matches with Level 5 (p.6). Marzano et al. (2008) pointed out that Analysis "involves reasoned extensions of knowledge" (p.6), i.e. matching, classifying, analysing errors, generating, and specifying. Level 3 Application in Bloom's Taxonomy was not made a distinct level in the New Taxonomy.*

superior to cognitive skills (including *Retrieval*, *Comprehension*, *Analysis* and *Knowledge Utilisation*) are added.

Besides the ranking of *Synthesis* and *Evaluation*[19], *Generalisation* is also a focal point of contention of taxonomies of educational objectives. Theorists classified it as a sub-skill under different levels in the hierarchies. Anderson & Krathwohl (2001) placed *Generalisation* under Level 2 *Understanding* as "conceptual knowledge", which "brings together large numbers of specific facts & events" and delineates "interrelationships among these specific details, classifications & categories" (p.52). Marzano et al. (2008) categorised *Generalisation* (which was described as "embedded in many components from Levels 4, 5 and 6 of Bloom's Taxonomy" (p.6)) as a higher level in the C*ognitive* domain of the taxonomy, Level 3 *Analysis*. *Generalisation* is also a skill in the Domain of *Knowledge* in the New Taxonomy, classified as a more complex skill than the use of *Vocabulary Terms*, *Facts* and *Time Sequences*, but less complex than the use of *Principles*.

The discrepancies of the rankings of cognitive skills, such as *Generalisation*, *Synthesis* and *Evaluation*, give grounds for research studies that gather empirical evidence from authentic performance of candidates to shed light on the levels of complexity of these skills and the use of cognitive models other than Bloom's Taxonomy (1956) in the validation process. In my study, the relative cognitive demands of these skills will be analysed with reference to the empirical evidence from the LS Examination.

One of the advantages of the New Taxonomy as compared to Bloom's classification is that it

[19] *Hauenstein (1998) believed that Generalisation is required for conceptualisation. Some assessment theorists positioned Generalisation at higher levels. For instance, Stobaugh (2014) introduced "inferring" as the skill in between "interpreting" and "applying", incorporating "generalisation from information" (p.22).*

integrates the metacognitive system into the cognitive domain, reflecting how values and goals underpin the performance of students. However, in Bloom's Taxonomy, values and goals were categorised as a separate domain, failing to draw their relationship with other skills. As the present study aims to further investigate how "values" were assessed and how the metacognitive system was applied in the DSE LS public examination, the New Taxonomy, in addition to Bloom's Taxonomy and the Revised Taxonomy (Anderson & Krathwohl, 2001), is an appropriate choice.

The large discrepancy in terminologies and hierarchical organisations of skills also entails divergence in the thresholds between lower- and higher-orders of (complex) thinking skills. Anderson & Krathwohl (2001) pointed out that skills in Levels 4, 5 and 6 of Bloom's Taxonomy are higher-order. Schraw & Robinson (2011) also concurred that skills resembling those in Levels 4, 5 and 6 of Bloom's Taxonomy, "Reasoning Skills", "Argumentation Skill", "Problem Solving & Critical Thinking[20]" and "Metacognition" (p.23), are core higher-order thinking skills. Corliss and Linn (2011) put forth a more specific classification of higher-order thinking skills for scientific enquiries (Table 2.7). Along a similar line as Bloom's Taxonomy, "demonstrating knowledge" is a lower-level skill, whereas "application" and "problem solving", which are similar to 4. *Analysis*, 5. *Synthesis* and 6. *Evaluation* in Bloom's Taxonomy, are in a higher order.

[20] *Bloom (1956) defined critical thinking as the intellectual abilities to generalise techniques, apply information (factual and/or theoretical), analyse the new situation in dealing with new problems.*

Table 2.7: Thinking skills for scientific enquiries (Corliss & Linn, 2011, p.221)

	Processes	**Skills Demonstrated in Assessments**
Lower-order	Demonstrating knowledge of concepts, laws, theory, procedures, and instruments	*Recall* *Define* *Describe* *List* *Identify*
Higher-order	Applying knowledge and procedures to solve complex problems	***Formulate questions*** *Hypothesise/ Predict* ***Design investigations*** *Use models* *Compare/ Contrast/ Classify* *Analyse* *Find solutions* *Interpret* *Integrate/ Synthesise* *Relate* *Evaluate*

In Dewey's (1937) words, an enquiry is a "transformation of a puzzling indeterminate situation into one that is sufficiently unified to warranted assertion" (as cited in Ormerod, 2006, p.900). "Constructing knowledge through issue-enquiry" is stipulated in the *Curriculum and Assessment Guide* for LS (The Curriculum Development Council (CDC) and HKEAA, 2014, p.18) as the key learning strategy. In this regard, learning LS involves strategies similar to learning through scientific enquiries as illustrated by Corliss and Linn (2011). To align with the enquiry-based nature of the subject, the definition of higher-order thinking skills for scientific enquiry, shown in Table 2.7, was adopted in this study. However, as the present study focuses on the written examination, "Formulate questions" and "Design investigation" (in bold in Table 2.7) which are assessed in the School-based Assessment only, were not the skills under investigation.

2.4.3 Cognitive models from learning sciences

Defining assessments as measurements of what students have learnt, the mental processes adopted in performing assessment tasks could be analysed with reference to cognitive models from learning sciences in the validation process. Bloom's Taxonomy, the Revised Taxonomy (Anderson & Krathwohl, 2001) and the New Taxonomy (Marzano et al., 2008) do not provide clues to the sequence of mental processes candidates adopt during assessments. Therefore, the learning theories and models that will be applied in the present research will be examined in this section.

The postulation of Posner et al. (1982) on knowledge construction in two distinctive stages, i.e. assimilation and accommodation, is an example of work along this line in learning sciences. In the 2000s, learning sciences of examinees, which focus on identifying and evaluating cognitive models as a measurement tool to assess achievement in learning and large-scale educational tests, began to capture the attention of educational researchers, for instance, Hanushek (2009), Leighton & Gierl (2007, 2011).

Leighton and Gierl (2011) evaluated some diagrammatic cognitive models in major academic learning domains (reading, science and mathematics) in terms of their applicability in large-scale assessments in the U.S. These models were chosen for review here as they have received substantial attention and empirical verification. Among the cognitive models reviewed by Leighton and Gierl (2011), the one on scientific enquiry is more relevant to the LS assessment, which also adopts an issue-enquiry approach.

The Scientific Discovery of Dual Search (SDDS) model of Klahr and Dunbar (1988) and Kuhn's (2001, 2005) Knowledge/Phases of Inquiry (KPI) model were reviewed by Leighton and Gierl (2011) as tools for the design and development of large-scale educational assessments. The use of

diagrammatic presentation to outline the key processes of scientific enquiries and the availability of theoretical and empirical support make these two models a viable basis for analysing the performance of thinking skills in LS examinations.

The SDDS model illustrated "the basic and most relevant knowledge and skills inherent to scientific reasoning and discovery" (as cited in Leighton & Gierl, 2011, p.123) This model stated that the problem solver generates a hypothesis with reference to his/her own knowledge, which will then be tested or evaluated for its adequacy with evidence from designed experiments. Klahr and Dunbar (1988) identified two types of methods to make progress from initial knowledge states to goal states – strong and weak methods (as cited in Leighton & Gierl, 2011). Strong methods refer to the use of domain-specific prior knowledge (for instance, the location of continents on a world map for the Geography candidates), whereas the weak methods involve domain-general knowledge, which is more useful in tackling novel problems. Examples of weak methods are: generate and test, hill climbing (the use of information "from intermediate steps to determine the most efficient way to proceed" (Klahr and Dunbar,1988, as cited in Leighton & Gierl, 2011, p.127)), means-ends analysis ("a comparison between states to determine whether a sub-goal should be introduced" (Klahr and Dunbar,1988, as cited in Leighton & Gierl, 2011, p.127)), planning (such as producing an outline), and analogy (making reference to an already-solved problem).

Kuhn's (2001, 2005) KPI model illustrated the cognitive processes of knowledge seeking[21] and delineated the strategies involved: enquiry (formulating an enquiry question), analysis and evidence evaluation skills, which are the core skills for an enquiry. The enquiry process is determined by the *Dispositions* underpinning the deployment of the strategies, the "procedural

[21] *This is also known as the model of knowing.*

and declarative *Meta-level* knowing" (Kuhn, 2005, p.140). The *Dispositions* may or may not be domain-specific strong methods in the SDDS. The incorporation of the *meta-level* processes crosses paths with the New Taxonomy (Marzano et al., 2008). *Meta-level* procedural knowing is the understanding of "when, where and why to use" (Leighton, 2011, p.164) strategies for gathering and interpreting evidence, for instance, "what it means to have adequate information on a topic, the differences between facts, opinions, and theories, and how evidence is coordinated to inform theories" (Leighton and Gierl, 2011, p.139). These domain-independent *Meta-level* strategies resemble some of the weak strategies put forth by Klahr and Dunbar (1988), such as means-end analysis, planning and analogy.

In the KPI model, Kuhn (2005) identified the procedural aspects of analysis, inference, enquiry[22] and reasoning; and ordered them from least to most effective (Table 2.8), shedding light on the possible criteria for differentiating the performance of candidates on an assessment task. From Table 2.8, the use and integration of evidence, and the consideration of alternatives/ counterargument/ counter-evidence are the key procedural aspects distinguishing the most effective deployment of these knowledge-seeking strategies, providing a more refined framework for analysing the mental processes for higher-order behaviours in taxonomies, such as Bloom's and the New Taxonomies. This concept of higher-order thinking skills is shared by Alexander (2011), who depicted these skills as complex, evidence-seeking, reflective, analytic and transformational. Along a similar vein, Corliss & Linn (2011) also considered complexity (which is effectiveness in Kuhn's terms in Table 2.8) as the key to distinguishing between higher- and lower-order thinking skill.

[22] *Leighton and Gierl (2011) referred to enquiring as "identifying questions, assumptions, or issues to investigate" (p.165).*

Table 2.8: Procedural aspects of Knowledge Seeking Strategies (Kuhn, 2001, p.2; Kuhn, 2005, p.85, 87, 89)[23]

	Analysis	**Inference**	**Enquiry**	**Reasoning**	
Least effective (Lower-order)	*Ignore evidence*	*Ignore evidence*	*Generate outcomes*	*No synthesis of evidence*	
	Interpret evidence to fit theory	*Represent theory without evidence*	*Generate best outcomes*	*Simple argument, narrative evidence*	
	Interpret as an illustration of theory	*Represent theory with selected or illustrative evidence*	*Examine variance in outcomes across distances*	*Simple corroboration of evidence, presenting counterarguments*	
	Compare selected instances as support for a theory	*Represent theory with supportive evidence only*	*Find out what makes a difference in outcome*		
Most effective (Higher-order)	*Choose for comparison two instances that allow the theory to be tested*	*Represent theory in relation to both supportive and unsupportive evidence*	*Find out if X makes a difference in outcome*	*Integration of supportive and other evidence*	*Discounting of alternative verdicts*
					Justification of alternative verdicts

According to Leighton and Gierl (2011), cognitive models can be evaluated for the purpose of educational assessment in terms of three criteria: grain size, measurability and instructional relevance. The first two are more relevant to this study on assessments.

Grain size refers to the depth and breadth of knowledge and skills that can be measured by using the model. The SDDS model describes the underlying cognitive processes in fine grain size. However, the detailed underlying processes do not map onto the interest of test developers, who aim at a coarse-level of knowledge and skills shown in the performance on assessment tasks. According to Leighton and Gierl (2007), the granularity of the KPI model is coarse as reflected

[23] *The procedural aspects of analysis, inference and enquiry were originated from Kuhn (2005), whereas those on reasoning were after Kuhn (2001).*

in the inclusion of values and the illustration of general strategies for a scientific enquiry.

The second evaluation criterion, measurability pertains to the feasibility of developing test items from the knowledge and skills illustrated in the model. In this dimension, even though both models have been validated with empirical data, they have not been adapted or translated for the purpose of developing items for large-scale assessments. Comparing the two models, the fine grain size in the SDDS model makes it more difficult to match with the educational objectives of large-scale assessments. On the other hand, experimental tasks, such as the evaluation of evidence (Kuhn 2001), have been designed to measure various aspects of the KPI model, demonstrating the research possibility in measuring cognitive processes by deploying the model.

With regard to the grain size and measurability, the coarse-grained KPI model is apparently more appropriate for analysing the performance of higher-order thinking skills of candidates in the LS examination. As for the SDSS model, instead of applying the whole model, the strong and weak strategies offer a plausible pathway to link knowledge and skill requirements for an assessment task. The applicability of the SDDS model was illustrated by the comments of Leighton and Gierl (2011) on the model as "a conceptual framework to be used as an overarching set of processes for how scientific reasoning and discovery occur within a variety of content domains of science" (p.217). Hence, in the current research, the strong and weak strategies and the differentiation of the effectiveness of thinking skills shown in Table 2.8 were some of the criteria for validating the examination of higher-order thinking skills in LS.

2.5 From Literature to the Present Study

In view of the dominance of quantitative, psychometric research on the validity of large-scale examinations, this study hopes to open a new avenue for research interests. L. S. Leung's (2017) research on Liberal Studies, which is by far of the closest relevance to the present study topic, was based on the perception of teachers and students on the wash-back effects of the public examinations. From a perspective different from L. S. Leung, the present research aims to devise an assessment validation process and illustrate the implementation and impacts by deploying actual performance of candidates in a public examination.

This study therefore focuses on a procedure for evaluating a large-scale examination based on quantitative and qualitative data, adopting the Argument-based Approach of Kane (2013, 2015). Empirical evidence of cognitive skills demonstrated by candidates who took the 2015 HKDSE LS Examination was analysed by employing cognitive models as the conceptual framework: Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001) and the New Taxonomy by Marzano et al. (2008) for evaluating the differentiation of *Knowledge*, *Information-handling, Synthesis and Evaluation* by the Levels of Performance of the examination; the New Taxonomy by Marzano et al. (2008) and Kuhn's (2001, 2005) KPI Model for analysing the sequential thinking processes and higher-order thinking skills of *Integration of Evidence* in *Argument*s and *Metacognition* (Table 2.9). The findings have been critically analysed to examine the applicability of the validation procedure on the evaluation of large-scale assessments. Although some cognitive models are empirically tested to specify the conceptual processes of enquiries, there is a lack of documentation on the application of learning scientific cognitive models, such as the KPI model, for test developers.

Table 2.9: The application of cognitive models in the validation process

Research Question	Thinking Processes to be Evaluated	Cognitive Model Applied
(2) To what extent is the substantive validity of the 2015 HKDSE examination justified? *(2a) Can the examination differentiate the Levels of Performance of candidates?*	• Differentiation of *Knowledge*, *Information-handling*, *Synthesis and Evaluation* by the examination	• Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001) and the New Taxonomy by Marzano et al. (2008),
(2b) Can the 2015 HKDSE LS Examination assess the higher-order thinking skills of candidates specified in the Level Descriptors?	• Sequential thinking processes • Higher-order thinking skills: Integration of Evidence in Arguments	• Kuhn's (2001, 2005) KPI Model
	• Higher-order thinking skills: Metacognitive skills	• New Taxonomy by Marzano et al. (2008),

How can these models be applied in the validation of the 2015 HKDSE LS examination? In what ways can the examination differentiate the levels of mastery of cognitive skills? This study aims to answer these questions, as well as conducting an in-depth analysis of the thinking processes of candidates who achieved various Levels of Performance in the examination. The limitations and merits of the procedure for evaluating a large-scale assessment, as exemplified by the validation of the 2015 HKDSE LS Examination, will be examined with a view to informing test developers of ways for improvement. The methodology of this study will be discussed in the following chapter.

CHAPTER 3 RESEARCH DESIGN AND METHODOLOGY

3.1 Aims and Research Questions

In view of the inclination of literature towards the theoretical issues of assessment validation rather than empirical research on methodology as pointed out by Deluca (2011), this research aims to devise a framework for evaluating the content and substantive validity of a large-scale examination. Messick's (1995) postulation of construct validity and Kane's Argument-based Approach (2013, 2015) are adopted as the conceptual framework to assessment validation. Empirical data of the 2015 HKDSE LS Examination from primary and secondary sources were deployed to illustrate the validation process probing into the content and substantive validity of the examination.

The research questions are as follows:

(1) *To what extent is the content validity of the 2015 HKDSE LS examination justified?*

The alignment of the Level Descriptors vis-à-vis the Assessment Objectives and the requirements of the 2015 examination was examined to evaluate the content validity.

(2) *To what extent is the substantive validity of the 2015 HKDSE LS examination justified?*

(2a) Can the examination differentiate the Levels of Performance of candidates?

The differentiation of various Levels of Performance of candidates was examined by applying cognitive models by Bloom (1956), Anderson & Krathwohl (2001) and Marzano et al. (2008).

(2b) Can the 2015 HKDSE LS Examination assess the higher-order thinking skills of candidates specified in the Level Descriptors?

Cognitive models by Marzano et al. (2008) and Kuhn (2001, 2005) were deployed for analysing the application of higher-order thinking skills by candidates. In addition, the

alignment of the thinking skills specified in the descriptors of higher Levels of Performance of the examination with the cognitive models was examined. Here, the aim is to contribute to the understanding of the cognitive demand of the HKDSE LS examination by deploying cognitive models.

3.2 Relevant Theoretical and Conceptual Ideas

Based on the unified approach to assessment validity postulated by Messick (1995), multiple sources of evidence were drawn for evaluating the content and substantive aspects of validity in this study.

The Argument-based Approach posited by Kane (2013 and 2015) formed the analytical framework for the evaluation of the content and substantive aspects of the 2015 HKDSE LS Examination (Diagram 3.1). With an aim of evaluating the validity as defined by *The Standards* (AERA et al., 2014), the appropriateness of "the interpretation and use" of the examination, the Validity Arguments of my study as developed in my dissertation proposal are as follows (T. Y. G.Leung, 2017):

> *(1) The assessment objectives and the assessment criteria of the 2015 HKDSE LS Examination align with the Level Descriptors;*
> *(2) The Level Descriptors appropriately differentiate the performance of candidates;*
> *(3) The 2015 LS Examination assesses the higher-order thinking skills of candidates specified in the Level Descriptors.*

Argument (1) was evaluated based on evidence for the content aspect of the examination, elicited from a content analysis of the Assessment Objectives, the assessment criteria of the examination questions and the Level Descriptors. To justify the substantive validity, Arguments (2) and (3) were verified with evidence from live scripts, nominal group discussions of examiners and a think-

aloud study[24] (which will be further explained in Session 3.4).

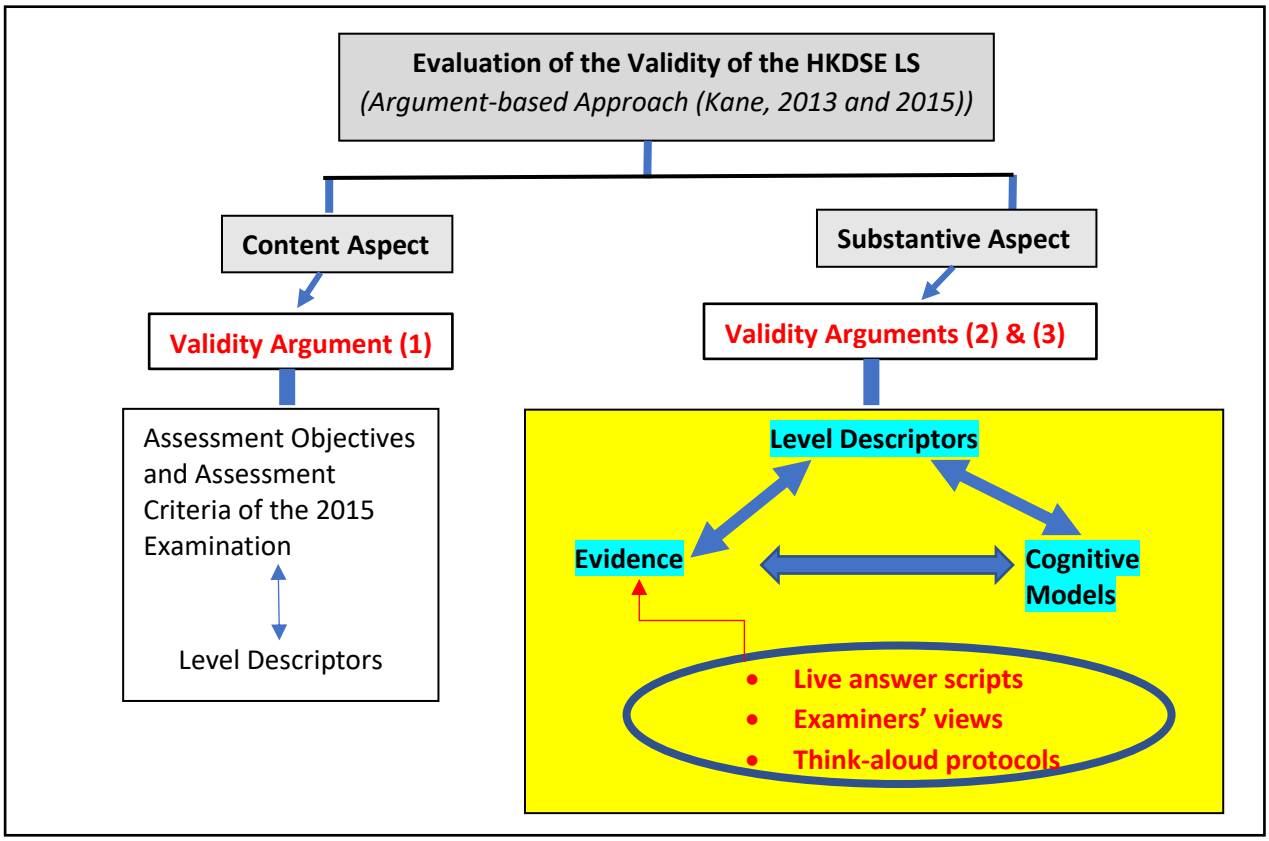
Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the New Taxonomy by Marzano et al. (2008) and Kuhn's (2001, 2005) KPI Model formed the conceptual framework for the analysis of the appropriateness referred to in the Validity Arguments (2) and (3): i.e. the differentiation of the thinking skills and the assessment of higher-level thinking process as stipulated by the Level Descriptors. Bloom's Taxonomy is chosen for my study as it is the origin of the development of cognitive models. As acclaimed by Schraw & Robinson (2011), it is as a model with "longevity", which "conceptualised both content and cognitive processes in a manner that spanned a broad spectrum of sophisticated skills" (p.12). The Revised Taxonomy, which supplements the classical Bloom's Taxonomy, strengthens the basis for the alignment analysis in the validation process.

In comparison with Bloom's Taxonomy, the more recent taxonomy by Marzano et al. (2008) postulated the relationship between knowledge/ concepts/ values and the cognitive processes, which encompass the metacognitive aspects. In view of how it supplements for the inadequacy identified in Bloom's Taxonomy, the New Taxonomy was also incorporated into the theoretical framework of the validation process to analyse qualitatively the cognitive skills of candidates, as exhibited in their answer scripts of the examination.

In addition, Kuhn's learning science model, KPI illuminates the procedural aspects of the cognitive processes of learners and so it is appropriate for the analytical framework of the sequence of higher-order thinking processes of candidates in the LS examination.

[24] *In my study, live examination scripts from candidates, who took the 2015 HKDSE LS Examination, were re-scored by examiners. Examiners discussed the scoring standards in a series of nominal group discussions. Among these candidates, 10 were selected for a think-aloud study of one of the questions in Paper 1.*

Diagram 3.1: The conceptual framework of the research (T. Y. G. Leung, 2017, p.36)



3.3 Research Design

3.3.1 Pragmatic Approach

My research adopts a pragmatic approach as the guiding principle. Morgan (2007) defined pragmatism as a research approach paying "equal attention to both the epistemological and technical 'warrants'" (p.68) by "a properly integrated methodology" (p.73). He also asserted that social science research adopting a pragmatic approach stresses "shared meanings and joint action", as opposed to being bound by "some external systems that will explain our beliefs to us" (p.66, 67). Owing to the inadequate literature on the methodology of assessment validation, which may constitute an "external system" for my research to be based on, a practicable validation process has been devised, incorporating "shared meanings and joint actions" of examiners.

Mixed methods were deployed in this research as a number of academics suggested the appropriateness of this methodology for research guided by pragmatism. Based on the ideas of Johnson and Onwuegbuzie (2006), Shannon-Baker (2016) put forward that "pragmatism is outcome-oriented and interested in determining the meaning of things" (p.322), without a pre-determined methodology or worldview, making mixed methodology a viable choice. Johnson and Onwuegbuzie (2004) believed that pragmatism "helps shed light on how research approaches can be mixed fruitfully" (as cited in Biddle & Schafft, 2015, p.323). As explicated by Biddle & Schafft (2015), researchers may encompass multiple methods with an aim of warranting "assertability of knowledge claims given available data, possibilities for analysis, and available resources" (p.323). Shannon-Baker (2016) pointed out the emphasis of pragmatic research approach on the "underlying belief of complementarity" (p.325) of quantitative and qualitative methodologies. Pragmatists, including Jackson (2000) have stated pragmatism as an approach that helps gain benefits from various methodologies (as cited in Ormerod, 2006). Along a similar vein, a mixed methods approach allows researchers to gather evidence from multiple strands of evidence for

assessment validation as advocated by Messick (1995), DeLuca (2011) and Cook et al. (2016). Adopting the pragmatic approach, the assessment validation process devised in this study sourced quantitative and qualitative evidence from primary and secondary sources (including a live script study, nominal group discussions and a think-aloud study), which will be further explained in the following sections.

Guided by the pragmatic approach, this research study does not aim to find the "truths" of the performance of all candidates in the 2015 HKDSE LS Examination, which is the concern of positivists. Pragmatism serves as a "bridge" (Shannon-Baker, 2016, p.325) between positivism and post-modernism. In the discussion of research paradigms by Usher (1996), positivists contended that research should be objective, while on the other end of the scale, post-modernists advocate subjectivity in the construction of views towards the world. The ontology of being objective has been challenged by various scholars. Habermas (n.d.) was one of them, who believed that objectivity in research is a matter of degree (as cited in Mitchell & Mitchell, 2003). Beista (2010) argued that "once-and-for-all truths" do not exist, instead "'knowledge' can only provide us with information about our actions and their results" (as cited in Shannon-Baker, 2016, p.325). Along this line of thought, my study assumed that the sequence of thinking processes may vary among individual candidates, rather than aiming to find out a "true" representation of the thinking processes of all candidates. Besides, the live script study and the think-aloud study provide secondary sources of evidence for the performance of thinking skills and the sequence of thinking processes deployed by the group of candidates in the study, rather than the general performance of the whole candidature of the examination. Unlike research adopting post-modernism, my study is evidence-based, instead of being absolutely subjective.

Oriented to the pragmatic approach, in my research, the design and the conduct of the assessment

validation process was underpinned by the principles of the pragmatic approach: abductive reasoning, intersubjectivity and transferability (Morgan, 2007, Shannon-Baker, 2016).

Abductive reasoning denotes a research process switching between induction and deduction. According to Biesta (2010), Dewey (1916 & 2009) and Maxcy (2003), the pursuit of knowledge by the pragmatic approach is a "constantly revised product of experience" (as cited in Biddle & Schafft, 2015, p.323).

Intersubjectivity entails reflexivity. Rather than taking either an absolutely subjective or an objective role, in Morgan's (2007) terms, a researcher adopting the intersubjective approach in the research process places emphasis on reflexivity and "a sufficient degree of mutual understanding" (p.72) with the participants, entailing processes for communication and persuasion in social research. In the words of Dewey (1925), pragmatists measure "existential reality" from a mixture of objective and subjective perspectives (as cited in Feilzer, 2010, p.8). Trainor & Graue (2014), though not referring to pragmatism specifically, contended that reflexivity is one of the criteria for a rigorous qualitative study.

Research underpinned by pragmatism aims at transferability rather than generalisability. Generalisability is a widely-discussed concern about research findings adopting the pragmatic approach (Feilzer, 2010). Unlike positivism, which aims to unveil the truth in reality, as postulated by Morgan (2007), the pragmatic approach presumes that the results of the research may be "usable in a new set of circumstances" (Morgan, 2007, p.72). This is referred to as the "transferability" of the research findings by Morgan.

Taking on a pragmatic approach, this study incorporated the interpretive enquiry framework for

assessment validation proposed by DeLuca (2011), accompanied by a quantitative analysis of scores for live scripts. Among the scarce literature on research approaches for the assessment validation process, DeLuca (2011) proposed an interpretive enquiry framework, which integrates the "dialectic, hermeneutic and transgressive forms of enquiry" "within current argument-based structures for the collection, analysis and representation of validity evidence" (p.303). The dialectic and hermeneutic modes tie in with the principles of the pragmatic approach: abductive reasoning and intersubjectivity. While the dialectic and the transgressive modes emphasise "reflexivity" (Moss, 1996 and Lynch, 2000 as cited in DeLuca, 2011, p.311), the hermeneutic mode "encourages conversations and interpretive work between and across stakeholder groups" (p.312).

Applying abductive reasoning in this research, the "product" of the research was "revised" by triangulating quantitative and qualitatively evidence. In the deductive process, cognitive models provided the basis for qualitatively and quantitatively analysing the empirical evidence sourced from live examination scripts, the group discussions and the think-aloud study. Subsequently, the data were deployed in an induction process to illuminate modifications to the cognitive models to represent the thinking skills of candidates deployed in an examination.

The "reflexive" orientation (Morgan, 2007, p.72) led this research to a reflection on the procedures to evaluate the validity of examinations. In DeLuca's (2011) terms, the transgressive mode of enquiry was adopted. The belief underpinning this mode is the transformative derivative of the pragmatic approach as discussed by Biddle and Schafft (2015), which focuses on the power imbalance in society. In the research world, the power imbalance exists with the dominance of quantitative research, which was considered "hegemony" by some academics. Shannon-Baker (2016) discussed the insufficiency of the "hegemony" of quantitative data by quoting Nightingale

(2003) from her research on Geography,

> *"Using different methods gives feminists an opportunity to demonstrate how hegemonic representations, such as remote sensed data for understanding land cover change, are insufficient."* *(p.331)*

As for assessment validation, a similar inclination to quantitative data can also be observed. Deluca (2011) commented that "contemporary measurement-based approaches" in assessment validity are based on quantitative "psychometric elements on student scores" (p.307). In this regard, with an aim of bringing about enhancement to the 2015 HKDSE LS Examination, another characteristic of the pragmatic approach – "creating practical solutions to social problems" (Shannon-Baker, 2016, p.325) is demonstrated.

Intersubjectivity, or in DeLuca's (2011) terms, the dialectic mode, was also the principle guiding the design of the validation process. The example of a dialectic enquiry of assessment validation provided by DeLuca (2011): the scoring by a group of stakeholders, resembles the live script study by examiners, which serves as the secondary data set in this research. The examiners built consensus and "shared" their views on the performance of candidates in the "joint action" (Morgan, 2007) of nominal group discussions. Instead of being absolutely objective, I scored the live scripts based on the "shared meaning" with the examiners on the same Scoring Grid.

Lastly, the transferability of the proposed assessment validation process will be considered in the Chapter 7. The implications: the factors contributing to, the limitations and the applicability of the proposed process to the validation of large-scale examinations will be discussed in Chapter 7.

3.3.2 Mixed Methods Approach

Guided by the pragmatic approach, a mixed method study, which was advocated by Creswell (2014) as an approach allowing "a stronger understanding of the problem" (p.215) with a variety of data collection methods, was adopted in this research. According to Messick (1995), multiple sources of evidence should be gathered for assessment validation. A number of academics also supported the use of mixed methods, including Ormerod (2006), Biddle and Schafft (2015) and Shannon-Baker (2016) as discussed in the previous section.

In response to the research question, both qualitative and quantitative data were collected from primary and secondary sources in my research for strengthening the evidential support. The primary and secondary data sources will be further explained in Section 3.4. Following the ideas of Kane and Messick, Cook et al. (2016) advocated a "methods neutral" approach (p.1359) to assessment validation, in which a multi-sourced data collection and analysis are determined by the purpose of the assessment. Koretz and Hamilton (2006) concluded from their research on the evaluation of assessment validity that evaluation frameworks relying "solely on psychometric evidence have been identified as problematic for use in contexts of high-stakes assessments" (as cited in DeLuca, 2011, p.307). Qualitative evidence will complement quantitative scores for a multi-faceted evaluation of assessment validity. Qualitative analysis, in Jick's (1979) terms, provided data for "between-method" triangulation with the quantitative data (as cited in Feilzer, 2010).

To illustrate a mixed-method assessment validation process, an evaluation of the content and substantive validity of the 2015 HKDSE LS Examination was conducted. Evidence was gathered from multiple sources as suggested by Shaw et al. (2012), including the authentic performance of candidates, views from stakeholders and think-aloud protocols. Shaw et al. (2012) believed that a

mixed-method approach can provide various strands of evidence for triangulation in the evaluation of the appropriateness of the interpretation and use of the examination results, which is the definition of assessment validation by *the Standards* (AERA et al., 2014).

Firstly, to evaluate the content validity of an examination, the alignment of Assessment Objectives and the assessment criteria for the examination with the Level Descriptors was analysed qualitatively.

In response to the second research question, an embedded explanatory sequential mixed approach described by Creswell (2014) was deployed to investigate some selected scripts of the examination both quantitatively and qualitatively. According to Creswell (2014), "perspectives of individuals" (p.231), which were in the form of the judgement of some experienced examiners in my study, were incorporated. Taking into account the inadequacies in the literature on the methodology of assessment validation, an empirical study on a large-scale public examination was conducted based on 144 live scripts[25]. A quantitative analysis on the scores awarded to these scripts according to a grid developed from the Level Descriptors was carried out. Subsequently, some of the scripts were selected for a further qualitative analysis according to the scores awarded and the nominal group discussions of experienced examiners. In this regard, some samples embedded in the quantitative study of the 144 answer scripts were analysed in a thematic analysis in a bid to delve into the details of the thinking skills of candidates. Both the convergent mixed approach and the exploratory sequential approach are not viable. In the former approach, conducting a statistical analysis and thematic analysis of the scripts in parallel makes it difficult to justify the criteria for selecting scripts for a thematic analysis. The exploratory sequential

[25] *Scripts from 72 candidates in the joint study, each with answers to 3 questions in Paper 1 and 1 question in Paper 2, together with 72 answers on each of the six questions in Papers 1 or 2 from the HKEAA Webpage were studied.*

approach, which starts off with a qualitative analysis of all the 144 answer scripts, is too time-consuming and ineffective in the selection of samples for a thematic analysis.

In a nutshell, my study comprises a quantitative analysis for portraying a statistically-tested overview of the differentiation power of the Levels of Performance by the examination. To elicit evidence on the details of the cognitive processes performed by candidates, the selected scripts and the think-aloud protocols were analysed qualitatively.

3.4 Data Collection

To illustrate an assessment validation process for a large-scale examination, primary and secondary data of the actual performance of candidates in the 2015 HKDSE LS Examination were deployed. The secondary data were taken from a joint study between the HKEAA and the HKUGA conducted in 2015-2016. Scores from four experienced examiners on 72 scripts, the nominal group discussion by examiners and the think-aloud protocols from the joint study were analysed quantitatively and qualitatively in the present study. To make up for the lack of scripts of Levels 1, 2 and 3 in the joint study, 72 scripts from the HKEAA website, which constituted the set of primary data, were scored by myself in accordance with the same Scoring Grid as in the joint study. Owing to the fact that only two answers of the candidates were scored by the examiners that in the joint study[26], the rest of the answer scripts were also scored to add to the set of primary data, with a view to obtaining a more comprehensive picture of the performance of each candidate in the examination. Approval for the use of the data from the joint study in my EdD study was granted by the HKAGE and the HKEAA (Appendix V). The collection of data in the joint study

[26] *Each candidate answered all the three questions in Paper 1 and one from the three questions in Paper 2. To enhance the representativeness of the scripts and the ease of scoring, the question in Paper 1 with a mark closest to that of the typical performance of the level attained by that candidate in the 2015 HKDSE was selected for the joint study.*

and the primary data will be discussed in this section. As the focus of my study is on the methodology for assessment validation, the details and the rationales of the data collection of the secondary data will also be explicated to justify the appropriateness of the data.

3.4.1 Secondary data

The joint study comprises a live script study, which scored two answers from each of the scripts of 72 candidates who took the 2015 HKDSE LS Examination, a nominal group discussion of four examiners and a think-aloud study of 10 candidates. All the participants in the joint study, including the examiners in the live script studies and the participants of the think-aloud study were informed that the data collected will be used for research studies by the HKAGE and me, as an examination officer of the HKEAA and an EdD candidate. The research purposes of the joint study and my own research were clearly communicated to the examiners and the think-aloud study participants as shown in the transcripts below and the email correspondences to the examiners (Box 3, Appendix V).

> *"The data collected will be used for research purposes by both the HKAGE and myself. The joint study between the HKAGE and the HKEAA aims at finding out the strengths and weaknesses of the members of the HKAGE in the 2015 HKDSE LS Examination. My doctoral degree research topic is "An Evaluation of the Validity of a Large-scale Assessment". The 2015 HKDSE LS Examination will be evaluated in terms of assessment validity and further enhancement of the development of question papers and the Level Descriptors will be derived.*
>
> *All data collected will be presented in an anonymously manner in reports."*
>
> *(An extract from the transcript of the Pre-meeting)(Translated from Cantonese)*

> *"The data collected will be used for research purposes by both the HKAGE and myself. The joint study between the HKAGE and the HKEAA aims at finding out strengths and weaknesses of the members of the HKAGE in the 2015 HKDSE LS Examination. My doctoral degree research aims at evaluating the examination for further enhancement.*
>
> *All data collected will be presented in an anonymously manner in reports."*
>
> *(An extract from the transcript of the introduction to the participants of the think-aloud study)(Translated from Cantonese)*

I took up the role of the designer of the data collection method in the joint study, including the development of the Scoring Grid and the think-aloud study, as well as the facilitator in the nominal group discussions and think-aloud study. To maintain the research validity in the data collection process, the potential conflict of interest in my dual role as the examination officer developing the LS examination and an evaluator of the assessment validity was minimised by having independent examiners in the nominal group to score the live scripts without knowledge of the actual Levels attained by the candidates. With a view to safeguarding the validity of the data and eliminating the potential bias of the test developer in the evaluation process, nominal group discussions were adopted. I was a facilitator, rather than a participant in the discussions. Examiners scored independently and then voiced their opinions on the scoring criteria before discussing for consensus. The same principle of research validity was adopted in the conduct of the think-aloud study. I was a demonstrator of the think-aloud process in answering another sub-question in the examination and a facilitator providing prompts for participants to continue in times of silence. In order not to create a tense situation by the presence of an examination officer, which deviates from the setting in a written examination, the participants were well-informed that this was not an oral examination. I also paid heed to withholding facial expressions or verbal hints on the answers in the think-aloud study.

In this thesis, the scores awarded by the examiners, the audio files of the discussions among

examiners (which were not analysed in the joint study) and the think-aloud study that were collected in the joint study constituted the secondary data sets and were analysed quantitatively and qualitatively, in accordance with my research objectives and analytical framework, which are different from those in the previous joint study.

Evidence from live scripts, think-aloud protocols or expert judgement is advocated by assessment scholars as a possible source of data for validation. Pellegrino et al. (2016) and Shaw et al. (2012) shared the views on the sources of evidence for assessment validation, including expert analyses of the cognitive requirements of test items, studies on "cognitive protocols" and "item and test performance" in relation to the demands on cognitive processes (Pellegrino et al., 2016, p.68), as well as teachers' surveys and classroom observation. As mentioned in the discussion on the research of L. S. Leung (2017) in Section 2.3, teachers' surveys and classroom observation were not deployed in my study because the focus of my research is on the actual performance of candidates in the live examinations rather than the teaching and learning process before the examinations.

3.4.1.1 Live Script Study

The joint study aimed to investigate the performance of 72 members[27] of the HKAGE in the 2015 examination by analysing the scores awarded by examiners according to a Scoring Grid derived from the Level Descriptors. As the majority of the members attained Levels 3 to 5**, the joint study excluded Levels 1 and 2. The examiners re-scored one question from each of Papers 1 and 2. The averages of the scores among the examiners were deployed to minimise subjectivity and enhance the reliability of the scores. Having independent scorers of the live scripts is a means to

[27] *Members of the HKAGE were gifted students nominated by secondary schools.*

reduce conflicts of interests that may arise due to my dual roles of the examination developer and the evaluator of the examination validity[28]. Owing to the deployment of the same scoring method for the primary and secondary data, the details will be provided in Section 3.4.2, which is on the sampling of primary data.

3.4.1.2 Nominal Group Technique

In the joint study, three nominal group discussions[29] were conducted, in which four experienced examiners in the examination "independently generated their ideas" (Van De Ven, & Delbecq, 1974, p.606), then presented their views, followed by a discussion for decision-making on the scores to the live scripts. Preceding the meetings, each examiner scored the scripts according to the grid in Appendix I. In the meetings, they discussed the rationales underpinning their scores and arrived at a consensus on scores with larger discrepancies. The scores and the views of the examiners made up the secondary data set for quantitative and qualitative analysis in this dissertation. The views of examiners provided a source of evidence for different levels of thinking skills displayed in the answer scripts in the qualitative analysis in my study.

A modified nominal group technique facilitated the consensus-building among the examiners on the skill requirements at different Levels of Performance. As put forward by Van De Ven, & Delbecq (1974), this technique facilitates the decision-making process in face-to-face meetings. In their words, individual members independently "generate their ideas on a problem or task in writing" (p.606). In the joint study, "their ideas" on the answer scripts were expressed in the form

[28] *The examiners were not full-time employees of the HKEAA and they participated in the marking and grading processes on a contract basis.*

[29] *The discussions were not analysed in the joint study. Since the meetings aim to discuss the scores of the live scripts, instead of asking the examiners to write down their ideas (as suggested in the nominal group technique), they were asked to input their scores to Excel Forms, which were compiled and presented to examiners in each meeting.*

of scores to the answer scripts according to the Scoring Grid (Appendix I). In the meetings, examiners presented their views on the performance without discussion. Subsequent to all examiners' presentations, they discussed and made final deliberation on the scores to the scripts.

Van De Van and Delpecq (1974) compared the nominal group technique, the Delphi method[30] and the conventional discussion group process. They concluded that the former two methods were more effective in giving rise to more ideas and perspectives for consideration and bringing about group satisfaction in the decision-making process. Nevertheless, in the Delphi method, different from the marking and grading process of LS, members are anonymous to one another and do not participate in physical meetings. To simulate a judging panel meeting in the grading process of LS, the nominal group technique, which involves a face-to-face decision-making process to achieve a collaborative decision on the scores, was adopted to elicit evidence of candidates' skill performance.

The foci for discussion and timeframe of the nominal group meetings are shown in Table 3.2.

Table 3.2: Nominal Group Meetings

	Focus for discussion	**Timeframe**
Pre-meeting	The scoring grid	2 hours
Meeting 1	The performance of Batch 1 (Level 5 and Level 3)	3 hours (one month after the pre-meeting)
Meeting 2	The performance of Batch 2 (Levels 5* and 5**)	3 hours (one month after Meeting 1)
Meeting 3	The performance of Batch 3 (Level 4)	3 hours (one month after Meeting 2)

In the pre-meeting, the facilitator briefed the examiners of the objectives of the study. To help the examiners to understand the skill and knowledge requirements as described on the Scoring Grid,

[30] *As defined by Van De Van and Delpecq (1974), the delphi method refers to "the systematic solicitation and collation judgements on a particular topic" through a set of "sequential questionnaires interspersed with summarised information and feedbacks' in writing" (p.606), without any face-to-face meetings.*

samples attaining Level 5 in the 2015 LS Examination were discussed. Level 5 samples were chosen for the pre-meeting due to previous experiences in grading meetings, which showed that it is easier for examiners to come to consensus on the top category of the Levels of Performance. Examiners were given one month before each meeting for scoring the scripts. To rule out the influence of the actual marks or Levels of Performance attained and to elicit judgement based solely on the actual performance and the skills specified in the Scoring Grid, blind marking was conducted. The Levels of Performance that each batch attained were revealed at the end of each nominal group discussion. Since the majority of the members of the HKAGE attained Level 3 or above, the secondary data from the joint study only comprised samples from these levels. For a more comprehensive analysis of all Levels of Performance, scripts of Levels 1 and 2 were taken from the HKEAA Website as primary data, which will be explained in Section 3.4.2.

The conduct of the pre-meeting can be justified by the "training group effect" as discussed by Baird et al. (2017). A "group culture" was found to be cultivated, which facilitated the building of consensus in the marking standard, by the training before the scoring process. The "group culture" will be further examined in Chapter 7 on the applicability of the validation process.

All the meetings took the following structure: (as first described in my EdD proposal (T. Y. G. Leung, 2017))

1. *Some scripts with similar scores and scripts with divergent scores were selected by the facilitator.*
2. *Each examiner explained their rationales for their scores.*
3. *The average scores were revealed.*
4. *Examiners discussed to come to a consensus on the scoring criteria.*
5. *The level attained by the scripts in the examination was revealed and the performance characteristics of the level were recapped.*

3.4.1.3 Think-aloud Study

The think-aloud study within the joint study provided a set of secondary data for a thematic analysis of the thinking processes employed by candidates in response to a question in the 2015 Examination. Think-aloud techniques are theoretically based on Vygotsky's (1962) concept of "inner speech" (as cited in Charters, 2003, p.69). According to Ericsson and Simon (1980), think-aloud methods allow probing into the verbal form of working memory (as cited in Charters, 2003). Olson et al. (1984) commended think-aloud as "effective ways to assess higher-level thinking processes" (as cited in Charters, 2003), which is Research Question 2(b) in my present study. Similarly, Gardner (2012) suggested that the cognitive procedures or strategies used to "construct meaning" and "develop mental models" (p.191) (which may be adopted for the fulfilment of the requirements of the questions in the examination in my study) could be elicited by think-aloud studies. Pellegrino et al. (2001, 2016), Shaw et al. (2012) and Gardner (2012) unequivocally pointed out that think-aloud studies can be a source of evidence for assessment validity. Gardner (2012) quoted the think-aloud study of Ericsson and Simon (1984) on the problem-solving processes of students, which are also the thinking processes required for the assessment of LS. Shaw & Imam (2013) proposed a triangulation of the findings from a think-aloud study with statistical analysis and text analysis of scripts in the evaluation of assessment validity. Therefore, think-aloud study provided evidence of the cognitive procedures adopted by candidates for a triangulation with the live script study in the evaluation of the validity of the LS examination.

The retrospective think-aloud study in the joint study, which formed a set of secondary sources for my current study, was conducted on ten candidates from the 900 members of HKAGE, about 5 months after they sat for 2015 HKDSE LS Examination. In the think-aloud study, participants were asked to verbalise their mental processes adopted in response to a data-response question (Paper 1, Question 3(b)). The sampling method will be explained in the Section 3.4.1.4.

Paper 1 Question 3(b) of the 2015 HKDSE LS Examination (HKEAA, 2015):

> *"(b) With reference to the sources provided, identify and explain two global concerns arising from the trends in international tourism you described in (a)."*

This question requires candidates to infer and justify global problems by using the data. According to Corliss and Linn (2011), inference and justification involve complex higher-order thinking skills, which are the focus of Research Question 2(b) in my study. The time concern is another factor for the choice of this question. Therefore, an answer that could be verbalised in about 15 to 20 minutes is not too long to deter participants from taking part in, making it suitable for the think-aloud study.

Before the participants worked on Question 3(b), the researcher demonstrated thinking aloud with Question 3(a). All cases of silence, which may be indicators of obstacles in the thinking processes, were noted down and prompts such as "what are you thinking about", were given.

Subsequently to "answering" the question, participants were asked to clarify some parts of the mental process, for instance, the reasons for stopping or hesitating, the sources of knowledge or concepts applied, the values they held in relation to the question and the development of the values. The use of retrospective questioning can help to clarify some of the incomplete data from the working memory and supplement the processes (Qi (1998) as cited in Charters, 2003 and Ericsson and Simon (1984) as cited in Nielsen et al., 2002). Nisbett and Wilson (1977) (as cited in Nielsen et al., 2002) suggested retrospective questioning help check against the possibility of modifying the thinking processes during verbal reporting. The whole think-aloud protocols, including the replies to the retrospective question will constitute the qualitative data sets in my current study.

As put forward by Someren et al. (1994), the generalisability of think-aloud protocols is a concern

for researchers. Individuals may not adopt the same cognitive strategies in approaching the same task in an examination and therefore, generalising cognitive strategies for all candidates from think-aloud studies is not possible. However, this study did not aim to generalise from the qualitative data. Instead, Morgan's (2007) idea of transferability is assumed. In other words, the findings of my study may only represent the cognitive strategies of some candidates attaining higher levels (the Levels attained by the participants of the think-aloud study will be discussed in the following section) in the examination and illustrate some possibilities of the procedural aspects of higher-order thinking. The transferability of the validation process with a think-aloud study will be discussed in Chapter 7.

3.4.1.4 Sampling of the secondary data

In this section, since one of the aims of this research is to devise an assessment validity evaluation process, the sampling of the secondary data will be detailed to justify the quality of the secondary data.

Firstly, for the selection of live scripts in the joint study with the HKAGE, a multi-stage sampling approach (Robson, 2011) was deployed. Samples were drawn from the 900 members of the HKAGE sitting for the 2015 LS Examination. To ensure a comparable number of samples from each Level of Performance, 72 candidates attaining Levels 3 to 5**, which were the levels attained by the majority of the members, were selected by stratified random sampling[31].

The number of scripts selected at various levels reflected the distribution of the 900 candidates.

[31] *The stratification was based on the distribution of the Levels of Performance obtained. A total of 72 candidates were selected in view of the budget, including the payment to the HKEAA for using the scripts and that to the four examiners, and the time frame for the study.*

Since Level 4 was the most commonly attained level by these candidates in the LS examination, the majority of samples were drawn at this level. Levels 1 and 2 were not incorporated in the joint study because of the relatively smaller proportion of members of the HKAGE attaining these levels.

Among these 72 scripts, 22.2% were answers in English and 77.8% in Chinese. The ratio of scripts in English to Chinese was quite similar to that in the whole candidature. In fact, the language version is not a variable in my study because the grading mechanism was irrespective of the language used. All candidates, no matter whether they took the examination in English or Chinese, were differentiated in Levels of Performance according to the same set of Level Descriptors. To enhance the reliability of the scoring of the live scripts in two languages, examiners with profound experiences in setting the marking standard across languages in LS were invited to participate in the study.

For the think-aloud study, ten members of the HKAGE were selected by stratified random sampling among the 900 members[32] who took part in the 2015 LS Examination. Since the aim of my study is to delineate the cognitive procedures of candidates in the application of higher-order thinking skills (Research Question 2(b)), a think-aloud study with most of the participants (five Level 4 and four Level 5 and above) selected among those attaining Level 4 or above in the examination will be appropriate as secondary data for analysis in the present study. The performance of the only candidate attaining Level 3 will be compared with others attaining higher Levels.

[32] *The participants were drawn from the 900 members rather than the 72 candidates selected for the live script study because the live scripts provided by the HKEAA were anonymous.*

In the literature of qualitative research studies, a great variety of sample sizes could be found. Leighton (2017) cited examples of think-aloud studies by Kucan (1993), Kucan & Beck (1997) and Goldman & Saul (1990) with sample sizes ranging from a few to 64 participants. She suggested that the sample size of think-aloud studies hinges on the research objectives. Morse (2000) also concurred that the sample size of qualitative research should be determined by the scope and nature of the topic, as well as the quality of data. In Morse's (2000) words, studies with an "obvious and clear" (p.4) objective could be conducted with a small sample size. The think-aloud study in the joint study fulfilled this requirement as the participants were instructed clearly to verbalise the thinking process when answering a question. The demonstration prior to the participants' think-aloud task and the prompts at the time of silence also helped to keep the participants "on target" (Morse, 2000, p.4). Furthermore, the post-interviews after the think-aloud task, according to Someren et al. (1994), elicited specific aspects in relation to the research objectives, thus enhancing the usefulness of the protocols from each participant and justifying the conduct of the study with a smaller sample size.

Participation in the think-aloud study was voluntary. Selected participants[33] were contacted over the phone to brief them on the purpose of the study and arrange for a face-to-face think-aloud study in the offices of HKAGE.

The transcripts of the nominal group discussions, comprising the views of examiners on the typical performance of various levels and the performance that they had difficulties in reaching agreements on the scores, made up another set of secondary data and were analysed thematically in this dissertation. As this dissertation focuses on the validation process, the selection of

[33] *13 members were contacted over the phone, 10 of them participated in the think-aloud study. Each participant was informed that a book coupon worth $50 dollars would be given as travelling allowance to encourage their participation.*

examiners for nominal group discussions will also be explained in further detail in the following paragraphs:

Four examiners were selected[34] among the 16 senior examiners who have participated in the grading and marking process of the HKDSE LS Examinations since the first examination in 2012. Their participation was voluntary. All of them had more than 10 years' teaching experiences in LS (including the AS Level LS) and more than five years' experience in leading the marking and grading process of the HKDSE LS Examination. In Hong Kong, there are only 16 senior examiners who are familiar with the level requirements as well as the marking and grading standards of the examination. In other words, these four examiners constituted quite a reasonable proportion (25%) of the 16-member expert group responsible for the grading process. The familiarity with the level requirements is of paramount importance in scoring the scripts in the study. Besides, the participation of the authentic examiners in the nominal group discussion allowed an investigation into the judgement-making processes in scoring the cognitive skills in this dissertation.

There was a methodological challenge in the validation process deploying nominal group discussions. Despite the experiences that the nominal group members possessed in marking and grading, they had not tried scoring performances with the Level Descriptors by skill domain. To familiarise them with the scoring process, a pre-meeting was conducted with some samples of answer scripts.

The validity of the secondary data set, which comprises the scores and nominal group discussions

[34] *Taking into consideration the time for discussion and the budget, four examiners were selected for participation in the live script study in the joint study. The HKAGE funded the live script study. The four examiners were paid for scoring the 75 scripts ($20 per answer script) and for attending the meetings ($750 for a meeting of 3 hours).*

by the examiners, could be justified by the selection of group members among the experienced examiners. The four examiners were familiar with the consensus marking process, which has been adopted since the times of Advanced-Supplementary Level LS[35]. They were able to narrow down the differences in the scoring standards through discussions of the rating criteria and the performance of individual scripts. By studying the consensus marking process in the discussions, the perception of examiners on the skill requirements for various Levels of Performance in the Diploma was examined, which provided a perspective for analysing the answer scripts qualitatively.

3.4.2 Primary data

To conduct a more comprehensive study in fulfilment of my research objectives, the part of the answer scripts not scored in the joint study[36] and the scripts from the HKEAA Website were further analysed in accordance with the Scoring Grid designed by me (Appendix I) in the joint study. The scores and the answer scripts constituted the primary data set for quantitative and qualitative analysis.

In the present research, to conjure up a more comprehensive picture of the performance of each candidate, I scored the answers for the remaining two answers in Paper 1 in each of the 72 scripts, which were not scored by the examiners in the joint study. As the joint study just included the performance from Levels 3 to 5** and the number of scripts attaining Level 3 was relatively small, answer scripts of Levels 1, 2 and 3 in the 2015 examination (Table 3.3) from the website of the HKEAA (2015b, October 30) were rated with the same Scoring Grid in order to incorporate all

[35] *The Advanced-Supplementary Level LS was an elective subject implemented from 1994 to 2013.*
[36] *In the joint study, the answers to one question in each of Paper 1 and Paper 2 of the scripts of the members of the HKAGE were scored by four examiners. I scored the remaining two answers in Paper 1 in my present study.*

Levels of Performance in the examination. As a facilitator of the nominal group discussions, I also aligned my marking standard with the four examiners.

Table 3.3: The composition of the scripts for the present study

Level of Performance attained	5**	5*	5	4	3	2	1	Total
Number of scripts from the joint study	8	12	14	28	10			72
Number of scripts from the HKEAA website					4 answer scripts for each question (a total of 24 answers)	4 answer scripts for each question (a total of 24 answers)	4 answer scripts for each question (a total of 24 answers)	72

All the scripts were re-scored according to a scoring grid of six domains of thinking skills (Appendix I) developed from the Level Descriptors (as exemplified in Table 3.4). According to the report of *the Development of the Draft Level Descriptors for the LS Subject of the HKDSE Examination* (HKEAA, 2007), the Level Descriptors were developed along the basic dimensions of skills and knowledge (the first column of Table 3.4) stipulated in the Assessment Objectives. With a view to conducting an in-depth analysis of the skills and knowledge demonstrated in the live scripts, the basic dimensions were further teased out. *"Formulation of viewpoints, opinions and suggestions"* was subcategorised into three aspects of skills: *"Synthesise"*, *"Evaluate"* and *"Appreciate cultures/ values/ views"*, constituting a scoring grid of eight domains (Appendix I).

Table 3.4: Level Descriptors of typical Level 5 candidates[37] (HKEAA, 2014)

Basic Dimensions	**Domain**	**Candidates at Level 5 typically:**
Knowledge and Understanding	*Understanding and application of relevant knowledge, key ideas and concepts of the subject*	● show comprehensive knowledge and understanding of the key ideas and concepts of the subject by applying relevant knowledge and concepts to a diverse range of complex issues in particular contexts
	Interpretation and analysis of the interdependence among personal, local, national and global issues	● interpret and analyse coherently the interdependence among personal, local, national and global issues from different perspectives
Generic Skills	*Handling of relevant information*	● identify relevant information, organise and analyse information from a diverse range of sources
	Formulation of viewpoints, opinions and suggestions	● evaluate various viewpoints and synthesise their own opinions and suggestions on the basis of logical arguments and sufficient examples ● demonstrating open-mindedness and tolerance towards a wide range of views and values
	Respect of Evidence	● solicit and conceptualise evidence and show respect for evidence
	Communication of Ideas	● communicate ideas in a concise, logical and systematic way

In the Scoring Grid (Appendix I), the performance of candidates was differentiated in eight domains, with respect to the range of knowledge and concepts, the range of perspectives/ cultures/ values/ viewpoints considered, the complexity of the skills, the logicality of viewpoints, the sufficiency of evidence, the organisation and presentation (as illustrated by the key descriptive terms in Table 3.5). The strength of performance decreases from the highest score of A to the lowest score of D or E on the Scoring Grid Descriptions. Whether the domain is differentiated on a four- or five-point scale and how the performance was described in the Scoring Grid was mainly

[37] *The Level Descriptors (Table 1.2) were taken from the official website of the HKEAA, whereas the domains (the second column of Table 3.4) were from my own analysis. The last bullet point of all levels on the Level Descriptors was not studied in this research because this is more comprehensively reflected by the School-based Assessment rather than the written examination.*

based on the wording of the Level Descriptors. Some of the descriptions on the Scoring Grid were constructed with reference to cognitive models. According to Anderson & Krathwohl (2001), "generalise" is more cognitively demanding than "analyse". Therefore, the highest performance for *handling of relevant information* (Grid Square 2A in Table 3.5) was described as being able to "generalise", whereas that of Grid Square 2B was "analyse". Besides, the description for Grid Square 4A: "*evaluate …based on clear criteria/ standards"* was also based on the definition of *Evaluate* of Anderson & Krathwohl (2001) (p.31).

Table 3.5 Description of candidates' performance on the Scoring Grid – Numbers (1-8) indicate the Domains, whereas letters (A-E) for the strength of performance

Domain of Skill	**Description of the mastery of the skill**				
Understanding and application of relevant knowledge, key ideas and concepts of the subject	comprehensive	broad	general	basic	elementary
	1A	1B	1C	1D	1E
Handling of relevant information	**generalise**	analyse	interpret	identify relevant information	Identify some basic and simple information
	2A	2B	2C	2D	2E
Interpretation and analysis of the interdependence among personal, local, national and global issues	coherently from different perspectives	from different perspectives	appropriately from different perspectives	briefly from some perspectives	identify simple relationships from a few perspectives
	3A	3B	3C	3D	3E
Formulation of viewpoints, opinions and suggestions	synthesise … on the basis of logical arguments	synthesise …with partly reasonable arguments	elaborate on … from the sources with partly reasonable arguments/ with simple elaboration	irrelevant … ungrounded arguments	
	4A	4B	4C	4D	
	evaluate …based on clear criteria/ standards	compare … without clear criteria/ standards	explain various viewpoints/ entities separately or explain by simple arguments	one-sided arguments… comparison/ evaluation/ assessment	
	5A	5B	5C	5D	
	show appreciation … towards a wide range of people/ incidents/ views / values in the formulation of arguments	consider particular cultures/ universal values / views/ values	show limited awareness of different cultures/ universal values, the concerns…	elaborate on their own views based on their own/ values/ cultures; without sound justification	
	6A	6B	6C	6D	
Respect for evidence	sufficient	identify some evidence	identify limited evidence	irrelevant … give little/ no evidence	
	7A	*7B*	*7C*	*7D*	
Communication of ideas	concisely, logically and systematically	logically and systematically	in an organized manner	simple ideas	Express simple ideas briefly
	8A	*8B*	*8C*	*8D*	*8E*

Since the descriptions were derived from the Level Descriptors, most of the domains were differentiated into 5 categories[38] of performance (from A to E). However, the differentiation of *"Formulation of viewpoints, opinions and suggestions"* and *"Respect for evidence"* was not clearly stated on the Level Descriptors from Levels 1 to 4 (Table 3.6). To *"*evaluate*"* was only performed by Level 5 candidates and "evidence" was not mentioned at Level 1. Therefore, Domains 4 to 7 were differentiated on four categories only. Failure to perform the skills in these domains was scored D (from Grid Squares 4D to 7D), which was described as being "irrelevant", giving "little/no evidence". Grid Squares B and C for these domains were based on the descriptors from Levels 4 to 2 on the Level Descriptors.

Table 3.6 The description of "Formulation of viewpoints, opinions and suggestions" and "Respect for evidence" on the Level Descriptors (extracted from Table 1.2)

Candidates at this level typically:

Level 4	• elaborate on various viewpoints and synthesise their own opinions and suggestions on the basis of logical arguments and some examples • solicit evidence and show respect for evidence, demonstrating open-mindedness and tolerance towards different views and values
Level 3	• elaborate on viewpoints and give their own opinions and suggestions supported by arguments and some examples • identify and show respect for evidence, demonstrating open-mindedness and tolerance towards different views
Level 2	• describe viewpoints and give their own opinions and suggestions supported by a few examples • identify evidence, demonstrate tolerance towards particular views
Level 1	• list viewpoints and give some opinions and suggestions • identify and describe related information from their own viewpoints

For the quantitative analysis of the live scripts, basically 5 to 1 points were allotted to Grid Squares A to E on the Scoring Grid respectively. However, for domains with 4 categories, scoring Grid Square A was allocated 5 points, the same as that for all other domains differentiated on a 5-point scale because the description of the highest performance (from Grid Squares 4A to 7A in Table 3.5) was derived from the descriptors which stipulate the requirements for attaining Level 5 on the Level Descriptors (Table 1.2). Grid Squares 4D to 7D were the lowest and so 1 point was

[38] *Levels 5* and 5** were graded statistically, with the top 10% and the next 30% of Level 5 candidates to be Levels 5** and 5* respectively. The Level Descriptors of Level 5 are applicable to all candidates attaining Levels 5, 5* and 5**. As the Scoring Grid was designed based on the Level Descriptors, A (5 points) is the highest on the scale.*

allocated. Grid Squares B and C (Grid Squares 4B to 7B and Grid Squares 4C to 7C) were given points by evenly dividing the scale into four segments with the minimum to be 1 point and the maximum 5 points (Table 3.7). In other words, C was made equivalent to 2.3 points (5/4+1) and B was allocated 3.6 points (5/4+2.3) on a 4-point scale.

Table 3.7: Point-allocation to the scoring grids

Domain	Grid				
	A	B	C	D	E
1	*5*	*4*	*3*	*2*	*1*
2	*5*	*4*	*3*	*2*	*1*
3	*5*	*4*	*3*	*2*	*1*
4	*5*	*3.6*	*2.3*	*1*	
5	*5*	*3.6*	*2.3*	*1*	
6	*5*	*3.6*	*2.3*	*1*	
7	*5*	*3.6*	*2.3*	*1*	
8	*5*	*4*	*3*	*2*	*1*

3.4.2.1 Sampling of the primary data

All the scripts of Levels 1, 2 and 3 in the 2015 LS Examination available on the HKEAA Website constituted the primary data set. For each level, there are two scripts in Chinese and two in English for each of the 6 questions in the examination (three questions in each Paper). Therefore, 24 scores for each of Levels 1, 2 and 3 (Table 3.3) were obtained. Lewin (2011) suggested that the sample size of studies which do not aim to make generalisation on the whole population may be less than 30. Since this dissertation adopts a pragmatic approach, a generalisation of the performance of the whole candidature is not destined and a larger sample size at each level is not necessary.

Purposive sampling (Flick, 2014) was deployed in the selection of live scripts for the qualitative thematic analysis. Scripts which were awarded similar scores by the examiners, showing consensus in the scoring standard, were selected. As guided by pragmatism, sampling in this research does not intend to generalise the performance of all candidates, rather it should be

“intentionally purposive” to “accurately reflect context-bound impressions that may transfer to new situations” (Cook et al., 2016, p.1367). A consensus in scores may suggest some typical examples of the performance of the candidates at a certain Level of Performance in the context of the 2015 LS Examination. Together with the samples of typical performance from the HKEAA Website, a comparison across Levels of Performance was facilitated. The limitations of this validation process with the purposive sampling will be further discussed in Chapter 7.

Besides some typical cases, the nominal group discussion focused purposively on “extreme or deviant cases” as termed by Flick (2014, p.175). Those with a larger discrepancy of scores were taken as the “deviant cases”. Both types of cases were discussed in the nominal group meetings, which provided qualitative evidence for the performance of candidates at various levels.

3.5 Data Analysis

The content analysis, quantitative analysis and the thematic qualitative analysis adopted by this dissertation will be explained in this section.

3.5.1 Content Analysis

A qualitative analysis of the alignment of the Assessment Objectives, the requirements of the 2015 Examination and the Level Descriptors was conducted.

Investigation into the agreement between the requirements of an examination and the Assessment Objectives stipulated in the curriculum was advocated by Webb (1997, 2007) (as cited in Wyse and Viger, 2011) for assessment validation. The Webb alignment method is a procedure used to examine the alignment of Assessment Objectives with knowledge and cognitive skills that candidates are expected to apply in an examination. Since the curriculum of LS is skill-based, rather than content specific, the content analysis in my study was centred around cognitive skills only.

Nevertheless, the alignment analysis alone cannot suffice as an assessment validation. As commented by Wyse and Viger (2011), the Webb alignment method cannot determine "how easy or difficult a test item is for students" (p. 188) and the actual cognitive processes adopted by students when responding to the question. This content analysis illuminates the alignment of the expected skills to be performed, from the perspective of test developers. To investigate how and whether the expected skills are performed in the assessment, data from the live script study and think-aloud study were necessary.

3.5.2 Quantitative Analysis of the Live Scripts

The answer scripts of the members of the HKAGE and scripts on the HKEAA Homepage were analysed. Answers to each question were scored and analysed according to the performance in the following skill domains (Table 3.8) on the Scoring Grid (Appendix I), derived from the Level Descriptors:

Table 3.8: Domains in the Scoring Grid

(1) understanding and application of relevant knowledge, key ideas and concepts of the subject;
(2) handling of relevant information;
(3) interpretation and analysis of the interdependence among personal, local, national and global issues;
(4) formulation of viewpoints, opinions and suggestions:
(4a) Synthesis
(4b) Evaluation
(4c) Cultures/Values
(5) respect for evidence;
(6) communication of ideas

In response to Research Question 2(a) *Can the examination differentiate the Levels of Performance of candidates?*, ANOVA was conducted[39] to find out whether there are significant differences and variability of (i) the scores for skills between Levels of Performance; (ii) the means of the aggregate scores (calculated by averaging the scores in the eight domains of each answer script) between Levels of Performance.

Furthermore, to find out whether the differentiation of the performance is in line with the cognitive models (Research Question 2(a)), the correlations between the skills in the eight domains and

[39] *The joint study conducted a different statistical analysis of the scores by the examiners.*

analyses of the variables derived from the dichotomised scores of pairings among Domains 2 *Information-handling*, 4 *Synthesis* and 5 *Evaluation* were conducted.

3.5.3 Qualitative Analysis of the Live Scripts, Think-aloud Protocols and the Nominal Group Discussions

A thematic analysis was conducted on the live scripts, think-aloud protocols and discussion transcripts. Based on the data reduction process suggested by Miles and Huberman (1994) (as cited in Punch, 2014), these qualitative data sets were coded by the cognitive skills in Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the New Taxonomy by Marzano et al. (2008) and Kuhn's verdict reasoning (2001) (Table 3.9). Though "Describing" and "Relating" (in bold in Table 3.9) were not specified in the three models, in the trial coding of live scripts, they were found to be frequently applied in the assessment tasks. These two skills were identified by Corliss & Linn (2011) as enquiry skills under the categories of "demonstrating knowledge of concepts, laws, theory, procedures, and instruments" and "applying knowledge and procedures to solve complex problems" respectively. However, in the LS Examination, candidates were not only asked to describe their own knowledge and concepts, but also to describe the patterns, phenomena or problems by using the data, especially in Paper 1. In other words, candidates had to articulate the findings of the data analysis. To reflect the authentic mental processes, "Describing" and "relating" were also added to the list of cognitive processes, under the category of *"Analysis"*.

For *Knowledge*, coding was done in accordance with the hierarchical categories of Knowledge of various complexity as stipulated in the Revised Taxonomy (Anderson & Krathwohl, 2001).

Subsequently, abstracting as suggested by Miles and Huberman (1994) (as cited in Punch, 2014) was conducted on the live scripts and think-aloud protocols to delineate the hierarchical and sequential relationships among the cognitive processes, namely *Retrieval, Understanding, Analysis, Evaluation, Creation, Knowledge utilisation, Metacognition* (Table 3.9). Through this process of abstraction, the substantive validity of the examination was scrutinised in terms of the alignment between the differences in performance of candidates attaining various Levels in the examination with cognitive models: by Bloom (1956), Anderson & Krathwohl (2001), Marzano et al. (2008) and Kuhn (2001, 2005), addressing Research Question 2(a) the differentiation of the levels of thinking skills; and 2(b) the assessment of higher-order thinking skills by the examination.

Table 3.9: Codes for data reduction

	Cognitive Processes		Knowledge
Codes for data Reduction (from Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the model by Marzano et al. (2008) and Kuhn's verdict reasoning (2001))	● Recognising ● Recalling	*Retrieval*	● Principles ● Generalisations ● Time sequences ● Facts ● Vocabulary items ● Macroprocedures ● Tactics ● Complex combination procedures ● Foundational procedures
	● Interpreting ● Exemplifying ● Explaining	*Understanding*	
	● Differentiating ● Organising ● Attributing ● Classifying ● Analysing errors ● Generalising ● Summarising ● Inferring ● Comparing ● **Describing** ● **Relating**	*Analysis*	
	● Critiquing	*Evaluation*	
	● Generating ● Representing the judgement criteria ● Citing pieces of evidence (without connection) ● Corroborating the evidence ● Integrating the evidence to the judgement	*Creation*	
	● Decision making ● Problem solving	*Knowledge utilisation*	
	● Specifying goals ● Monitoring process ● Monitoring accuracy	*Metacognition*	

As well as delving into the differential skills of candidates attaining various levels in the examination in response to Research Question 2(a), the higher-order thinking skills displayed by the Level 4 and above candidates in the think-aloud protocols were also to be analysed (Research Question 2(b)).

For analysing the think-aloud protocols, "task analysis" suggested by Someren et al. (1994) was conducted. By adopting Kuhn's KPI model as the "normative model" (Someren et al., 1994, p.37) for abstraction, the cognitive processes of candidates were studied for constructing procedural models for the illustration of how higher-order thinking skills were deployed in the LS examination. The relationships among the *Meta-level* strategies, *Dispositions*, *Information-handling* and *Argument Formulation* in the KPI model were delineated by using empirical data from the think-aloud study.

The actual quantitative and thematic analysis of the primary and secondary data for the evaluation of the content and substantive validity of the 2015 HKDSE LS Examination will be discussed in the Chapters 4, 5 and 6.

CHAPTER 4 EVALUATION OF THE CONTENT VALIDITY OF THE 2015 HKDSE LS EXAMINATION

In this Chapter, an evaluation process of the content aspect of the validity of the 2015 HKDSE LS Examination will be conducted.

As defined by Pellegrino & Wilson (2015), an assessment should be in "parallel" with the curriculum (p.264). Following this line of thought, content validity can be justified if the Assessment Objectives delineated in the curriculum are in "parallel" with the requirements of the examination as shown in question papers and the Level Descriptors. To justify the content validity of the 2015 LS Examination, a content analysis was conducted to investigate the alignment among the Assessment Objectives, the requirements of the examination papers and the Level Descriptors.

Firstly, the six domains of the Level Descriptors[40], *Understanding and application of relevant knowledge, key ideas and concepts of the subject; Interpretation and analysis of the interdependence among personal, local, national and global issues; Handling of relevant information; Formulation of viewpoints, opinions and suggestions; Respect of Evidence* and *Communication of Ideas*, (in bold in Table 4.1) cover explicitly the Assessment Objectives (Table 1.1), except *e*, *k*, *n* and *o*.

> *"• (e) to recognise the influence of **personal and social values** in analysing contemporary issues of human concern;*
> *• (k) to self-manage and reflect upon the implementation of successive stages of the enquiry learning process in terms of time, resources and attainment of the objectives of the enquiry;*
> *• (n) to demonstrate an understanding and appreciation of different **cultures and universal values**; and*
> *• (o) to demonstrate **empathy** in the handling of different issues."*

[40] *Assessment Objective (k) can be fulfilled by the School-based Assessment in the form of an Independent Enquiry Study (IES), rather than the written examination. Therefore, it is out of the scope of this study.*

There may not be a one-to-one or word-for-word correspondence between the Assessment Objectives and the domains in the Level Descriptors. To begin with, Assessment Objective *h* fuses two domains: *handling of relevant information* and *formulation of viewpoints, opinions and suggestions.*

> *"(h) to **analyse issues** (including their moral and social implications), **solve problems, make sound judgments and conclusions and provide suggestions, using multiple perspectives**, creativity and appropriate thinking skills;"*

As for the affective elements, including values and empathy in the Assessment Objectives *e*, *n* and *o* above, the requirement for the performance in this aspect can only be traced at Level 5. Candidates attaining this level are described as being able to "demonstrate open-mindedness and tolerance towards a wide range of views and values" (Table 3.4) in the formulation of arguments. Therefore, even though affective elements, like values, are not explicitly described at levels other than Level 5, the performance criterion for the affective elements in these three Assessment Objectives could be manifested in the descriptions of the *formulation of viewpoints, opinions and suggestions,* as phrased in the Level Descriptors. The alignment between the Level Descriptors and the Assessment Objectives is shown in Table 4.1. The Level Descriptors on the Domain of *Knowledge and Understanding* cover Assessment Objectives *a*, *b*, *c*, *d* and *i*, whereas the Domain of *Generic Skills* covers Assessment Objectives *f*, *g*, *h*, *j*, *l* and *m*.

Table 4.1: The alignment between the Level Descriptors and the Assessment Objectives

Level Descriptor		**Assessment Objective** (the letters indicate the order in the Curriculum and Assessment Guide, abstracted in Table 1.1)
Basic Dimension	**Domain**	
Knowledge and Understanding	*Understanding and application of relevant knowledge, key ideas and concepts of the subject*	• *(a) to demonstrate a sound understanding of the key ideas, concepts and terminologies of the subject* • *(b) to make conceptual observations from information resulting from enquiry into issues* • *(c) to apply relevant knowledge and concepts to contemporary issues*
	Interpretation and analysis of the interdependence among personal, local, national and global issues	• *(d) to identify and analyse the interconnectedness and interdependence amongst personal, local, national, global and environmental contexts* • *(i) to interpret information from different perspectives*
Generic Skills	*Handling of relevant information*	• *(g) to discern views, attitudes and values stated or implied in any given factual information* • *(h) to analyse issues (including their moral and social implications)* • *(m) to gather, handle and analyse data and draw conclusions in ways that facilitate the attainment of the objectives of the enquiry*
	Formulation of viewpoints, opinions and suggestions	• ***(e) to recognise the influence of personal and social values in analysing contemporary issues of human concern*** • ***(n) to demonstrate an understanding and appreciation of different cultures and universal values*** • ***(o) to demonstrate empathy in the handling of different issues*** • *(h) to solve problems, make sound judgments and conclusions and provide suggestions, using multiple perspectives,* ***creativity*** *and appropriate thinking skills;* • *(j) to consider and comment on different viewpoints in their handling of different issues*
	Respect of Evidence	• *(f) to draw critically upon their own experience and their encounters within the community, and with the environment and technology*
	Communication of Ideas	• *(l) to communicate clearly and accurately in a concise, logical, systematic and relevant way*

Similarly, although "creativity" in Assessment Objective *h* is not mentioned explicitly in the Level Descriptors, Level 5 candidates are expected to be able to "synthesise their own opinions", which was taken as equivalent to "create their own opinions" by Anderson & Krathwohl (2001). Therefore, "creativity" is exhibited in the domain: *formulation of viewpoints, opinions and suggestions,* even though the exact wording is not employed in the Level Descriptors.

Direct reference is also not given to "evidence" on the list of Assessment Objectives. Albeit a Level 5 candidate is described as being able to "*solicit and conceptualise evidence and show respect for evidence"* (Table 3.4) in the Level Descriptors, this requirement is not clearly discerned by the Assessment Objectives. The Assessment Objective that matches closest to the domain on evidence is *f* (Table 4.1).

> *"(f) to draw critically upon their own experience and their encounters within the community, and with the environment and technology;"*

Cultural consideration specified in the Assessment Objectives is the only attribute that cannot be matched with the Level Descriptors. The appreciation of different cultures mentioned in Assessment Objective *n* cannot be traced in the way that Level Descriptors are phrased. There is a lack of description of the consideration of cultural aspects in candidates' performance in the examination.

From the above content analysis, most of the Assessment Objectives are explicitly described or embedded in the description of some other attributes of performance by the Level Descriptors, with the exception of "cultural consideration". To investigate whether the expected requirements of the examination as delineated by the Assessment Objectives are reflected in the 2015 Examination, an analysis of the alignment between the Level Descriptors and the question-specific requirements was conducted (Table 4.2).

Referring to Table 4.2, all the eight domains of the Level Descriptors were encompassed in the 2015 LS Examination. However, it is not expected and impracticable that each question should assess all the domains. While Domain 1 *Understanding and application of relevant knowledge, key ideas and concepts of the subject,* Domain 7 *Respect for evidence and* Domain 8 *Communication of ideas* were assessed in all questions, the other domains might not be covered

by all questions. For Domain 3 *Interpretation and analysis of the interdependence among personal, local, national and global issues*, even though the consideration of issues from multiple perspectives was required, the perspectives to be taken into account varied among the six questions. The two papers together allowed candidates to consider multiple perspectives: personal, local, national, global, environmental and economic perspectives.

Besides, since Paper 2 comprises extended-response questions, *Handling of given information* (Domain 2) was not required. In other words, this skill was only demonstrated in Paper 1 (Table 4.3). *"Synthesis"* and *"Evaluation"* were also not skills required for both papers. Paper 2, which comprises optional questions, was designed to provide equal opportunities for candidates to perform the thinking skills stipulated in the Assessment Objectives. Therefore, all the questions in Paper 2 required candidate to "synthesise" and "evaluate". According to the Revised Taxonomy (Anderson & Krathwohl, 2001), "synthesise" and "evaluate" are cognitive skills in the highest order and therefore Paper 2 consisted of more demanding questions. Questions assessing skills in the lower order of the Revised Taxonomy, i.e. "analyse", "describe" and "infer", which are described under *Handling of relevant information* (Domain 2) in the Level Descriptors, were found in Paper 1.

Table 4.2: The alignment between the Level Descriptors and the question-specific requirements of the 2015 LS Examination

Domain of Skill[#]	Paper 1			Paper 2		
	Question 1*	**Question 2**	**Question 3**	**Question 1**	**Question 2**	**Question 3**
1. Understanding and application of relevant knowledge, key ideas and concepts of the subject	✓	✓	✓	✓	✓	✓
2. Handling of relevant information	Analyse , Generalise the changes, Describe, Infer	Generalise the reasons	Generalise the trends, describe, infer	Nil	Nil	Nil
3. Interpretation and analysis of the interdependence among personal, local, national and global issues	National, Social	Personal, Local	Global, Social, Environmental, Economical	Local, Social	Local, Social, Personal	Global, Personal, Social
*4, 5, 6.** Formulation of viewpoints, opinions and suggestions*	Suggest measures; Evaluate effectiveness	Synthesise arguments; Evaluate impacts	Synthesise arguments, Evaluate impacts	Synthesise arguments, Evaluate effectiveness	Synthesise arguments, Evaluate impacts	Synthesise arguments, Evaluate effectiveness
7. Respect for evidence	✓	✓	✓	✓	✓	✓
8. Communication of ideas	✓	✓	✓	✓	✓	✓

Notes:

[#]The alignment with Domain 1 was indicated by ✓because there are numerous knowledge or concepts that candidates may use in their answers. The alignment with Domains 7 and 8 was also indicated by ✓because the domain names clearly denote the skill requirements.

**The questions are shown in Table 4.3.*

***On the Scoring Grid in Appendix I, Domain 'Formulation of viewpoints, opinions and suggestions' was further divided into 3 categories: 4 Synthesis, 5 Evaluation and 6 Cultures/Values.*

Table 4.3: The questions of the 2015 LS Examination[41] (HKEAA, 2015)

Paper 1		
Question 1	(a)	With reference to Sources A, B and C, describe the changes in the condition of *sannong* (agriculture, rural areas and farmers) in China. (5 marks)
	(b)	With reference to the sources provided, explain ***two*** social problems that might arise from the changes in the condition of *sannong* in China. (6 marks)
	(c)	For ***each*** social problem you identified in (b), suggest and explain ***one*** measure that could deal with it. Explain your answer with reference to the sources provided and your own knowledge. (6 marks)
Question 2	(a)	From Source A, identify and explain ***two*** reasons why an increasing number of young people in Hong Kong are undergoing plastic surgery. (6 marks)
	(b)	With reference to the sources provided and your own knowledge, should the Hong Kong government ban 'medically unnecessary' plastic surgery on under-18s through legislation? Justify your stance. (8 marks)
Question 3	(a)	Describe the trends in international tourism shown in Source A and suggest ***one*** potential benefit that might arise from the trends. Explain your answer. (4 marks)
	(b)	With reference to the sources provided, identify and explain ***two*** global concerns arising from the trends in international tourism you described in (a). (8 marks)
Paper 2		
Question 1	(a)	What factors do you think might influence press freedom in Hong Kong? Explain your answer. (8 marks)
	(b)	'A high degree of press freedom would enhance the effectiveness of governance by the Hong Kong government.' To what extent do you agree with this view? Explain your answer. (12 marks)
Question 2	(a)	What would you consider to be the barriers to achieving consensus among major stakeholders on the issue of standard working hours in Hong Kong? Explain your answer. (8 marks)
	(b)	Source A claims that "standard working hours are essential to the improvement of the quality of life of Hong Kong people". To what extent do you agree with this claim? Explain your answer. (12 marks)
Question 3	(a)	Explain the effects that the entertainment industry, as a global culture, may have on its audiences. (8 marks)
	(b)	'Soft power is the most effective way for governments to increase their influence in the world.' Do you agree with this view? Explain your answer. (12 marks)

[41] *The sources in Paper 1 and the stimulus materials in Paper 2 are not shown here.*

Even though *Respect for evidence* (Domain 5) is not explicitly delineated by the Assessment Objectives, it is specified in the question-specific Marking Guidelines as the assessment requirements, aligning with the Level Descriptors. As specified in the Marking Guidelines for Paper 1 Question 2 and Paper 2 Question 1[42], candidates were expected to use the sources or other examples in support of their viewpoints (Table 4.4). The requirement for using evidence was more explicit in Paper 1 Question 2(b). To score the top marking range, candidates had to "draw appropriately upon the relevant evidence from the sources and his/her own knowledge" (in bold in Table 4.4). The deployment of evidence is necessary in answering all the questions in Paper 2. For instance, in Paper 2 Question 1(b), candidates were required to deploy "relevant and valid examples/ observations in Hong Kong" (in bold in Table 4.4), which serve as supporting evidence to justify their arguments.

[42] *Paper 1 Question 2(b) was the only sub-question in Paper 1 requiring candidates to express their views. In Paper 2, all sub-questions are assessing candidates' ability to support their arguments by examples or their own observations.*

Table 4.4: Extracts from the Marking Guidelines of Paper 1 Question 2(b) and Paper 2 Question 1b (HKEAA, 2015)

Paper 1 Question 2(b)

Suggested Marking Guidelines	Marks
The candidate:	
… ● explains and justifies clearly and logically the extent to which he/she agrees that the government should ban 'medically unnecessary' plastic surgery on under-18s through legislation in view of the current situation of Hong Kong; **draws appropriately upon the relevant evidence from the sources and his/her own knowledge;** … *Points in support of the ban: explains clearly and in detail his/her arguments with the points of relevance in the sources and relevant and valid examples; …*	7-8

Paper 2 Question 1(b)

Suggested Marking Guidelines		Marks
The candidate:		
… *Points of agreement:*	*(A high degree of press freedom would enhance the effectiveness of governance by the Hong Kong government)* ***explains clearly and in detail his/her arguments with relevant and valid examples/ observations of Hong Kong**; …*	10-12

The content alignment among the Assessment Objectives, question-specific requirements and the Level Descriptors indicated the intended demands of the examination. As a majority of the Assessment Objectives (with an exception of cultural considerations) and the question-specific requirements are in line with the Level Descriptors, the requirements of the 2015 LS Examination comply with those stipulated in the Curriculum and Assessment Guide in general, demonstrating the content validity of the examination. In other words, the examination was designed to assess and differentiate the performance of candidates in terms of how well they fulfil the expected requirements as specified by the Assessment Objectives. However, does the examination demand more than what is specified in the Assessment Objectives with regard to "respect for evidence"? Is cultural consideration neglected in the examination, deviating from the Assessment Objective?

Does the examination allow candidates to demonstrate the fulfilment of the intended requirements and provide appropriate evidence for "inferences about score meaning or interpretation" (Messick, 1995, p.5)? To answer these questions, in addition to the content analysis of the fulfilment of the intended objectives of the examination, an empirical study on the substantive validity in the following chapter is indispensable.

CHAPTER 5 EVALUATION OF THE SUBSTANTIVE VALIDITY OF THE 2015 HKDSE LS EXAMINATION

In this Chapter, an evaluation process of the substantive aspect of validity will be illustrated by empirical data from the 2015 HKDSE LS Examination, using both quantitative and qualitative evidence from a live script study and nominal group discussions.

Following the ideas of Messick (1995) and Kane (2006), an assessment can be claimed to be valid if the "interpretation and use" of the assessment results is appropriate. In the case of the LS Examination, its validity lies in the appropriateness in (i) differentiating the performance of candidates into levels as stipulated in the Level Descriptors and (ii) in assessing higher-order thinking skills as set forth in the Assessment Objectives. Drawing from the analysis in the previous chapter, the Level Descriptors are largely in line with the Assessment Objectives (with the exception of the description of the considerations of cultures and values, as well as evidence). Therefore, a scoring system for live scripts in the examination based on the Level Descriptors can provide data for a quantitative analysis of the differentiation between levels, as well as reflecting the performance of candidates with regard to the fulfilment of the Assessment Objectives.

Adopting Kane's (2013, 2015) Argument-based Approach to evaluating the appropriateness of the examination, the differentiation power and assessment of higher-order thinking skills were investigated via an analysis of the alignment with cognitive models by Bloom (1956), Anderson & Krathwohl (2001) and Marzano et al. (2008).

The differentiation of the Levels of Performance by the examination will be analysed from two perspectives: the performance by skill domain and the overall performance of candidates.

5.1 The Differentiation of Performance by Skill Domain

To examine whether the 2015 LS Examination was capable of differentiating the Levels of Performance of candidates, one-way ANOVA was conducted to test for significant differences in the scores[43] awarded to the live scripts of candidates attaining different levels in the Examination, according to a Scoring Grid on a scale ranging from 1 to 5 points for each of the eight skill domains of the Level Descriptors (Tables 5.1 and 5.2).

Owing to the differences in the nature of the data collected by the joint study and that from the HKEAA Homepage, with the former being randomly sampled among the 900 members of the HKAGE, whereas the samples from the latter demonstrating typical performance of each question, the set of data from the joint study will be analysed on its own and subsequently, typical samples will be added for analysing a full spectrum of the Levels of Performance.

For the scripts in the joint study, for each domain, a one-way ANOVA (at a significance level of 0.05) showed statistically significant differences in the scores among candidates attaining Levels of Performance 3, 4 or 5 (p-values were under 0.001 for all domains (Table AII-1A-2, Appendix II)). However, post hoc comparisons using the Tukey HSD test (Table AII-1A-3, Appendix II) indicated that statistically significant differences in the scores for *Cultures/Values* and *Evidence* were not shown between Levels 3 and 4 although the means were in the expected direction (in bold in Table 5.1). The p-values of *Cultures/Values* and *Evidence* between Levels 3 and 4 were 0.330 and 0.063 respectively. All other domains had statistically significant differences between all pairs of Levels of Performance.

[43] *The average of the scores awarded by the 4 examiners for a skill domain in the joint study was taken as the score for a certain answer in the quantitative analysis. The other answers, not studied in the joint study, were scored by me and so there was only a single score.*

Table 5.1 The means and S.D. of scores by skill domain for answer scripts from the joint study[44] (from Table AII-1A-1, Appendix II)

	Level of Performance attained					
	5		4		3	
Skill Domain	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
1. Understanding	4.459	0.534	3.761	0.607	3.288	0.704
2. Information-handling	4.711	0.430	4.196	0.634	3.733	0.713
3. Perspectives	4.423	0.519	3.808	0.601	3.156	0.722
4. Synthesis	4.251	0.647	3.487	0.562	3.049	0.691
5. Evaluation	3.667	0.704	2.944	0.715	2.517	0.764
6. Cultures/values	3.717	0.681	**2.996**	0.670	**2.810**	0.718
7. Evidence	4.257	0.626	**3.368**	0.627	**3.097**	0.779
8. Communication	4.610	0.429	3.973	0.461	3.500	0.610
Number of Scores	**136**		**112**		**40**	

Table 5.2 The means and S.D. of scores by skill domain for answer scripts from the joint study and the HKEAA Homepage (from Table AII-1B-1, Appendix II) (the figures in bold and italics show the incorporation of by-domain scores for scripts from the HKEAA Homepage)

	Level of Performance attained									
	5		4		3		2		1	
Skill Domain	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
1. Understanding	4.459	0.534	3.761	0.607	***3.180***	***0.571***	***2.000***	***0.000***	***1.000***	***0.000***
2. Information-handling	4.711	0.430	4.196	0.634	***3.690***	***0.660***	***2.833***	***0.577***	***1.917***	***0.900***
3. Perspectives	4.423	0.519	3.808	0.601	***3.129***	***0.594***	***2.000***	***0.000***	***1.000***	***0.000***
4. Synthesis	4.251	0.647	3.487	0.562	***3.093***	***0.664***	***2.000***	***0.000***	***1.000***	***0.000***
5. Evaluation	3.667	0.704	2.944	0.715	***2.558***	***0.695***	***1.379***	***0.604***	***1.000***	***0.000***
6. Cultures/Values	3.717	0.681	2.996	0.670	***2.549***	***0.827***	***1.295***	***0.558***	***1.000***	***0.000***
7. Evidence	4.257	0.626	3.368	0.627	***2.900***	***0.842***	***1.758***	***0.655***	***1.000***	***0.000***
8. Communication	4.610	0.429	3.973	0.461	***3.317***	***0.541***	***2.000***	***1.000***	***1.000***	***0.000***
Number of Scores	**136**		**112**		***64***		***24***		***24***	

The ANOVA results changed with the addition of scores for the typical performance of Level 3 candidates on the HKEAA Homepage. Statistically significant differences were now indicated in the scores for skill domains between the 5 Levels in all but 2 cases (Table 5.2 and Table AII-1B-2, Appendix II). For Levels 1 and 2, significant differences were found except in the scores for

[44] *Each answer script from the joint study consists of answers to 4 questions and therefore, 4 scores were awarded to each domain. (Levels 5 and above: 34X4=136; Level 4: 28X4=112; Level 3: 10X4=40)*

Evaluation and *Cultures/Values*. (*Evaluation*: p-value = 0.297; *Cultures/Values*: p-value = 0.609 at a significance level of 0.05) (Table AII-1B-3, Appendix II). The candidates attaining these two levels performed close to the poorest category on the Scoring Grid (means of the scores were 1.000) for these two skill domains. As described on the Scoring Grid (Appendix I), the performance of these two skills scoring 1 point is as follows: "provided one-sided arguments/ described one of the entities/ pros/cons without comparison" and were able to "elaborate on their own views based on their own values/ cultures; without sound justifications".

The now significant differences in the ANOVA results of *Cultures/Values* and *Evidence* between Levels 3 and 4 may be explained by the incorporation of the typical performance at Level 3, which pulled down the scores for this category of these two domains (as shown in the means of these two skills in Tables 5.1 and 5.2), making them significantly lower than those for Level 4. The limitations stemming from the use of scripts of different nature will be discussed in Chapter 7.

The lack of statistically significant differences in *Evaluation* between Levels 1 and 2 may not lead straight to the conclusion that the examination failed to differentiate between candidates' performance on this aspect at these levels. In fact, *Evaluation* was not differentiated explicitly by the Level Descriptors of Levels 1 and 2 (HKEAA, 2014). At these two levels in the Level Descriptors, the bullet points pertaining to the domain on the formulation of viewpoints (which comprises the skills to *synthesise* and to *evaluate*) just describe candidates' consideration of "views", without mentioning the performance on "evaluation":

> *Level 2: "…demonstrate tolerance towards particular views"*
> *Level 1: "…identify and describe related information form their own viewpoints"*

Candidates attaining these two overall Levels for the subject are expected to "formulate viewpoints, arguments, opinions and suggestions" with the consideration of "particular views" or

"their own viewpoints", but not to *Evaluate,* Therefore, the lack of differentiation in *Evaluation* between Levels 1 and 2 as found in the ANOVA does not deviate from the performance as stipulated in the Level Descriptors.

However, for the domain of *Cultures/Values,* the same argument by the Level Descriptors failed to provide a sound explanation. As shown in the previous chapter, the description on cultural considerations has been missing throughout the five levels on the Level Descriptors, not only at these two levels.

Skills *4. Synthesis*, *5. Evaluation* and *6. Cultures/Values* are the Sub-domains comprising *"formulation of viewpoints, arguments, opinions and suggestions".* To examine the performance as a whole on the domain of *"formulation of viewpoints, arguments, opinions and suggestions"*, the average scores from its three skill sub-domains were computed and then a One-way ANOVA analysis was used for this derived domain. As significant differences were shown in the average scores of these 3 sub-domains (p-value<0.001) (Table 5.3 and Table AII-4B-3, Appendix II), the domain of *"formulation of viewpoints, arguments, opinions and suggestions"* on the Scoring Grid was differentiated between all levels.

Table 5.3: The means and S.D. of the average scores of Skill Domains: 4. Synthesis, 5. Evaluation and 6. Cultures/Values from Levels 5 to 1 (from Table AII-4B-1)

	Level of Performance attained									
	5		**4**		**3**		**2**		**1**	
	Mean	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
	3.876	0.544	3.142	0.553	2.731	0.566	1.686	0.336	1.00	0.00
No. of answer scripts	136		112		64		24		24	

Another observation from the One-way ANOVA on the scores for each skill domain at different Levels of Performance (Tables 5.2 and AII-1B-1, Appendix II) is that the mean scores for

Evaluation and *Cultures/Values* were far below the maximum (5 points) among the Level 5 candidates (*Evaluation*: (M=3.667) and *Cultures/Values*: (M=3.717)), attaining merely B on the Scoring Grid (3.6 points). Referring to the overall performance in *"formulation of viewpoints, arguments, opinions and suggestions"* (Table 5.3), which could be shown by the average of the scores for sub-domains: *4. Synthesis, 5. Evaluation* and *6. Cultures/Values*, the mean at Level 5 was also far below 5 (3.876). Not all candidates at the top level were able to evaluate "based on clear criteria/standards" and consider a range of cultures / values / views "in the formulation of arguments" as described on the Scoring Grid and the Level Descriptors for Level 5. The downward squeezed scales for these two domains might have led to the insignificant differentiation between Levels 1 and 2.

From the nominal group discussion, the examiners suggested that the lower scores for *Cultures/Values* could be explained by the question requirements of the 2015 Examination. One of the examiners, E2 pointed out Paper 1 Question 1 (Table 4.3) as an example not requiring a discussion of values:

> *E2: "This is not essay-writing. All questions have a scope…. Unlike P1Q3 and P2Q3, some questions (e.g. P1Q1) involve factual knowledge and concepts, e.g. sannong. It is not easy to show the values of the candidates."*
>
> *(Nominal Group Discussion 3)*

As shown in the content analysis (Tables 4.2 and 4.3), the consideration of various cultures/values/views was not explicitly required in all questions. For Paper 1 Question 1 (Table 4.3), it is legitimate for candidates to answer the question without discussing much about social disparity as a social problem in part (b) or exploring the changes in the life of people in different social groups after the implementation of the measures candidates suggested in part (c). Therefore, they might just consider a particular social group in this question and the scores for this domain in this question might be lower, pulling down the mean scores.

However, this does not imply that the content validity is undermined. Providing opportunities for candidates to analyse and make judgement on contemporary issues is the key to maintaining the "liberal nature of the subject" (HKEAA, 2014, p.131), which does not target assessing candidates' ability to give model answers. In other words, adhering to the curriculum, candidates may answer the questions from various perspectives and via different approaches, even in data-response questions. As shown previously in Table 4.2, since all the Assessment Objectives were covered when both papers were taken into account, a content-wise alignment with the Assessment Objective can be justified.

To verify whether the performance in some questions significantly pulled down the scores for Domain 6, an ANOVA was conducted on the by-question scores for this domain (Tables 5.4, AII-5-1, AII-5-2 and AII-5-3, Appendix II) at each Level of Performance. It was found that there was no statistically significant difference in the scores for Domain 6 among different questions in the examination (p-value=0.132). Examiners might have adjusted the scoring standard across the questions. A few scripts of Paper 1 Question 1, which did not indicate any consideration of *Cultures/Values*, were not scored for this domain because this was not a compulsory skill to be performed in the question. Even though the scores for Domain 6 were not significantly pulled down by a certain question, the possibility of the optional application of this skill in the whole examination leading to lower scores among the 8 domains cannot be dismissed. Some scripts might be scored lower for Domain 6 due to the lack of variety of *Cultures/Values* considered as mentioned in the nominal group discussion.

Table 5.4: The means and S.D. for the by-question scores for Domain 6 *Cultures/Values*

	Paper 1 Question 1		Paper 1 Question 2		Paper 1 Question 3		Paper 2	
	Mean	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
	2.915	0.967	3.215	1.099	2.909	1.010	2.879	1.108
No. of answer scripts	76		81		81		107	

Nevertheless, a similar argument of question requirements does not work in providing a sufficient explanation for the performance on *Evaluation*. All questions required candidates to put forth some criteria for assessment or judgement-making, demonstrating the skill of *Evaluation*. For Paper 1 (Table 4.3), candidates had to explain the effectiveness of the measures they suggest in Question 1(c), justify their own views in Question 2(b) and assess the extent and seriousness of the impact which deserves global concerns in Question 3(b). All questions in Paper 2 expected candidates to justify their viewpoints on some contemporary issues. Therefore, the lower scores for *Evaluation* even at Level 5 (M=3.667, SD=0.704) (Table 5.2), cannot be explained by the requirements of the questions.

In Section 5.3, the relatively poorer performance on *Cultures/Values* and *Evaluation* will be further analysed by cognitive models. To examine the performance on *Evaluation* and *Cultures/Values* between Levels 1 and 2, direct evidence other than the scores of the scripts by skill domain is necessary for a more in-depth analysis of the evaluation skills and the consideration of various views/ cultures/ values in the formulation of viewpoints in the live performance in an examination. In this regard, a qualitative analysis on the live scripts was conducted and will be discussed in Section 5.3.

5.2 The Differentiation of the Overall Performance

With reference to the views of Messick (1995) and Kane (2006), the validity of an examination hinges on the appropriateness in the "interpretation and use" of the assessment results. Hence, the validity of the 2015 LS Examination can be demonstrated in its appropriate differentiation of candidates' holistic performance by level, which was stipulated as the way of "interpreting and using" the assessment results on the Level Descriptors.

The overall averages of the scores in the eight domains were computed for each answer script[45] and analysed using a one-way ANOVA analysis (Table 5.5 and Tables AII-1B-1, AII-1B-2 and AII-1B-3, Appendix II) (at a significance level of 0.05) to examine the differentiation of the holistic performance in the examination. The ANOVA analysis of the data from the joint study yielded the same results (Table 5.6 and Tables AII-1A-1, AII-1A-2 and AII-1A-3, Appendix II) as that after the addition of scripts from the HKEAA Homepage. With the Scoring Grid developed from the Level Descriptors, the statistical analysis of the average scores provided evidence for cross-checking the appropriateness of the grading of the overall performance of candidates.

Table 5.5: The means and S.D. of the average scores for each answer script at Levels 5 to 1 (from Table AII-1B-1, Appendix II)

	Level of Performance attained									
	5		**4**		**3**		**2**		**1**	
	Mean	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
	4.242	0.436	3.551	0.465	3.022	0.495	1.894	0.199	1.059	0.101
No. of answer scripts	136		112		64		24		24	

[45] *An answer script refers to the answer for one question. Each candidate has to answer a total of 4 questions (3 in Paper 1 and 1 in Paper 2). From the joint study, the distribution of answer scripts at different levels was as follows: L5: 34X4=136; L4: 28X4=112; L3:10X4=40. From the HKEAA homepage, 4 scripts from different candidates were provided for each question, making up a total of 24 answer scripts for each of the Levels 1 to 3.*

Table 5.6: The means and S.D. of the average scores for each answer script at Levels 5 to 3 in the joint study only[46] (from Table AII-2-1, Appendix II)

	Level of Performance attained					
	5		4		3	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
	4.242	0.436	3.551	0.465	3.122	0.575
Number of candidates	136		112		40	

In view of the overall performance as shown in the live scripts, the lack of a statistically significant difference in the scores of individual domains, *Evaluation* and *Cultures/Values*, between Levels 1 and 2 discussed in Section 5.1 (Table AII-1B-3, Appendix II) did not undermine the overall differentiation power of the examination as stipulated by the Level Descriptors, which are holistic in nature.

The overall performance of candidates (Tables 5.5 and 5.6) attaining a specific level was significantly better than that of candidates attaining the subsequent lower level. As the scores were awarded according to a Scoring Grid derived from the Level Descriptors, the ANOVA results show evidence for the differentiation power of the requirements of the examination as stipulated by the Level Descriptors from Levels 5 to 1.

The relatively smaller standard deviation of the scores for Level 1 candidates (SD=0.101) in comparison with that of Level 5 candidates (SD=0.436) (Table 5.5) denoted that the performance of candidates at Level 1 was less varied. Another observation is that the median of Level 5 candidates was 4.278 (Table 5.7). As shown in Table 5.8, only 3.676% of candidates attaining Level 5 were awarded 5 points for all questions. The majority scored on average between 4 and 4.5.

[46] *Scripts for Levels 4 and 5 were all from the Joint Study and so the findings are the same as that in Table 4.9.*

Table 5.7: The variability of the average scores of answers attaining Levels 3, 4 and 5 in the joint study[47] (from Tables AII-2-1, AII-2-2 and AII-2-3, Appendix II)

		Level of Performance		
		5	**4**	**3**
Mean		4.242	3.551	3.122
Std. Error of Mean		0.037	0.044	0.091
Std. Deviation		0.436	0.465	0.575
Variance		0.190	0.216	0.331
Skewness		-0.392	-0.589	-0.277
Std. Error of Skewness		0.208	0.228	0.374
Percentile	25	3.940	3.320	2.732
	50	4.278	3.592	3.225
	75	4.649	3.862	3.556
Number of scores		136	112	40

Table 5.8: The distribution of the average scores of answer scripts in different score ranges at Levels 5 to 3 in the joint study

	Level of Performance attained								
	5			**4**			**3**		
Scores	**=5**	**<5 and >/=4.5**	**<4.5**	**>/=4**	**<4 and >/=3.5**	**<3.5**	**>/=3**	**<3 and >/=2.5**	**<2.5**
% of scripts	3.676	36.206	60.118	14.286	44.643	41.071	55.000	30.000	15.000
Number of scripts	**136**			**112**			**40**		

Since the descriptions of the performance of candidates scoring different points on the Scoring Grid were derived from the Level Descriptors, taking Level 5 as an example, candidates attaining this level were expected to perform as stipulated in the Level Descriptors and thus scoring 5 points. As such, a mean score of 4.242 (Table 5.7) reflected that some L5 candidates did not fully fulfil the requirements of all the eight domains stipulated by the Level Descriptors.

Notwithstanding the squeezed range of scores and the variability of the scores at higher levels, the One-way ANOVA of the overall mean scores for the script for each candidate reflected the

[47] *Since there were 4 examiners scoring the scripts in the joint study, the variability in the scores was analysed for the answer scripts from the joint study only, excluding those from the HKEAA Website.*

appropriateness of the Level Descriptors, based on which the Scoring Grid was designed, in distinguishing the overall Levels of Performance.

5.3 The Alignment with Cognitive Models

In this section, whether the examination differentiated the performance of candidates in the study in agreement with cognitive models, namely Bloom's Taxonomy (1956), the revised Bloom's Taxonomy (Anderson & Krathwohl, 2001) and the New Taxonomy (Marzano et al., 2008), will be discussed with reference to the quantitative and the qualitative data from the live script study.

Four to five samples[48] for each of the five Levels of Performance were selected for a thematic analysis of the performance in Domains 1 to 6 to investigate the alignment of the examination with the cognitive models. For Levels 1 and 2, since all the samples from the HKEAA Homepage illustrated the typical performance, they were all analysed thematically. As for the samples from the joint study, scripts awarded similar scores by the examiners were chosen, assuming that they might exhibit some performance characteristics the examiners concurred with. From each of the five levels, findings from one English script[49] were tabulated in Sections 5.3.1.2 and 5.3.2.2 to illustrate the performance in relation to the skill domains of *Knowledge*, *Information-handling*, *Synthesis* and *Evaluation*. English scripts were selected for tabulation so that authentic excerpts can be quoted without a loss of information in translation. To minimise the variables involved in the comparison, all samples that had answered Question 1 in Paper 2, which has an explicit

[48] *The number of scripts selected for the thematic analysis was determined by the availability of scripts. For each of Levels 1 and 2, there were four samples on the HKEAA Homepage. On the other hand, samples of Levels 3 to 5 were selected for the thematic analysis from those awarded similar scores by the four examiners. 5 samples were selected in Levels 4 and 5, while 4 samples in Level 3, which has a smaller total number of scripts in the joint study.*
[49] *Candidates have the option of taking the examination either in Chinese or English.*

requirement for a discussion on "values", "press freedom", were chosen.

> *Paper 2 Question 1*
> *"(a) What factors do you think might influence press freedom in Hong Kong? Explain your answer. (8 marks)*
> *(b) 'A high degree of press freedom would enhance the effectiveness of governance by the Hong Kong government.' To what extent do you agree with this view? Explain your answer." (12 marks)*

According to Newton et al. (2014) and Pellegrino et al. (2001), the validity of examinations can be studied in relation to the construct of the assessment. Bloom's Taxonomy (1956), the revised Bloom's Taxonomy (Anderson & Krathwohl, 2001) and the New Taxonomy (Marzano et al., 2008) were taken as the basis for analysis because all these taxonomies purport the relative orders of cognitive demands of thinking skills, though the order and the terminology may vary. To align with the taxonomies, the examination should differentiate the Levels of Performance of candidates in terms of the command of thinking skills. Candidates attaining a higher Level of Performance in the examination should be able to show a better mastery of the thinking skills at a higher rank in the taxonomies.

Evidence and *Communication* (Domains 7 and 8) will not be analysed by taxonomies of thinking skills in this section. This is because firstly, the use of evidence and communication skills is not a distinct category of cognitive skills in Bloom's Taxonomy, the Revised Taxonomy or the New Taxonomy. The exclusion of Domain 8 *Communication* in the following discussion can also be justified by the high correlations (Table 5.9) between the scores for this domain and all the other 7 domains, implying that the skill can be observed in the performance on all other skills in a written examination. Domain 7 will be analysed in Chapter 6 with reference to the think-aloud protocols and Kuhn's (2001, 2005) KPI model which stipulate the use of evidence for *Argument Formulation.*

The interplay between *Evidence*, *Knowledge*, *Perspectives*, *Cultures/Values* and other higher-order cognitive skills will be further discussed in Chapter 6 with findings from the qualitative analysis of the live scripts and think-aloud protocols, based on the KPI model (Kuhn, 2001, 2005).

Table 5.9: Correlations among skill domains (significant at $p < 0.01$)

Domains		1	2	3	4	5	6	7	8
1. Understanding	Pearson Correlation	1	0.794	0.946	0.890	0.816	0.830	0.863	0.929
	Number	360	252	360	360	358	345	360	359
2. Information-handling	Pearson Correlation	0.794	1	0.789	0.763	0.628	0.619	0.775	0.826
	Number	252	252	252	252	250	238	252	252
3. Perspectives	Pearson Correlation	0.946	0.789	1	0.902	0.810	0.801	0.865	0.928
	Number	360	252	360	360	358	345	360	359
4. Synthesis	Pearson Correlation	0.890	0.763	0.902	1	0.814	0.768	0.870	0.888
	Number	360	252	360	360	358	345	360	359
5. Evaluation	Pearson Correlation	0.816	0.628	0.810	0.814	1	0.781	0.779	0.800
	Number	358	250	358	358	358	344	358	357
6. Cultures/Values	Pearson Correlation	0.830	0.619	0.801	0.768	0.781	1	0.788	0.813
	Number	345	238	345	345	344	345	345	344
7. Evidence	Pearson Correlation	0.863	0.775	0.865	0.870	0.779	0.788	1	0.873
	Number	360	252	360	360	358	345	360	359
8. Communication	Pearson Correlation	0.929	0.826	0.928	0.888	0.800	0.813	0.873	1
	Number	359	252	359	359	357	344	359	359

5.3.1 The Knowledge Domain

With regard to the New Taxonomy (Marzano et al., 2008), *Knowledge* belongs to a discreet hierarchy separated from the "levels of processing" (p.2). The meta-analysis of Kreitzer and Madaus (1994) (as cited in Anderson & Krathwohl, 2001) provided evidence for this distinct category. Categorised in a similar manner, Anderson et al. (1994) put forth the Affective Domain, which includes *Cultures/Values*, as a distinct taxonomy from cognitive skills. Kuhn (2001, 2005) suggested that *Knowledge* and *Values* be grouped under *Dispositions*. Therefore, the performance

on *Knowledge* is not only shown in Domains 1 *Understanding* and 3 *Perspectives*, but also in Domain 6 *Culture/Values* on the Scoring Grid. Candidates have to deploy knowledge from various perspectives, such as the social, environmental or technological aspects, in answering questions, as well as knowledge pertaining to various cultures/values. The high correlations between Domains 1, 3 and 6 (Pearson Correlations: 0.946 between Domains 1 and 3; 0.830 between Domains 1 and 6 and 0.801 between Domains 3 and 6 respectively) (Table 5.9) provided another piece of evidence for the close relationship in the performance in these three skill domains. Hence, to examine the appropriateness of the examination with regard to the *Knowledge* domain, whether the differentiation of performance in Domains 1, 3 and 6 between levels goes in line with the hierarchy of *Knowledge* in the cognitive taxonomies has to be analysed.

5.3.1.1 Quantitative analysis

As discussed in Section 5.1, the ANOVA results for both Domains 1 and 3 (Tables 5.10 and AII-1B-2, AII-1B-3, Appendix II) (at a significance level of 0.05) showed statistically significant differences in the scores between the five Levels of Performance. It is evident that the examination differentiated the performance of candidates on these two skills. However, for Domain 6, statistically significant differences were found among Levels 2 to 5.

Table 5.10 The means and S.D. of scores for Domains 1, 3 and 6 for answer scripts attaining Levels 5 to 1 (from Table AII-1B-1, Appendix II)

	Level of Performance attained									
	5		**4**		**3**		**2**		**1**	
Skill Domain	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
1. Understanding	4.459	0.534	3.761	0.607	3.180	0.571	2.000	0.000	1.000	0.000
3. Perspectives	4.423	0.519	3.808	0.601	3.129	0.594	2.000	0.000	1.000	0.000
6. Cultures/values	3.717	0.681	2.996	0.670	2.549	0.827	1.295	0.558	1.000	0.000
Number of Scores	**132**		**116**		**64**		**24**		**24**	

In this section, whether the differentiation aligned with the cognitive models will be first examined

with reference to the design of the Scoring Grid (Appendix I). Under the presumption that examiners scored the scripts by adhering to the descriptions on the grid, the scores can be interpreted in terms of the performance descriptions on the grid, which was derived from the Level Descriptors. As such, the alignment of the differentiation by the Level Descriptors with the cognitive models can be examined.

For both Domains 1 and 3, the criterion on the Scoring Grid for differentiating the performance of candidates is the one suggested by the New Taxonomy: the complexity of "system of thoughts" (Marzano, 2008, p.3). Domain 1 was scored according to the comprehensiveness of "knowledge and understanding of key ideas and concepts" (Scoring Grid on Appendix I). Candidates scored higher for the domain if they were able to show an understanding of comprehensive knowledge, which demands a more complex "system of thoughts" for organising more knowledge and concepts. For Domain 3, the differentiation of performance as described on the Scoring Grid hinged upon the variety of perspectives taken into consideration by a candidate. Integrating a greater variety of perspectives in the *Knowledge Utilisation* process involves more complex "mental procedures" as termed by Marzano et al. (2008, p.3).

Referring to Table 5.10, the mean scores for Level 5 scripts for Domains 1 (M=4.459) and 3 (M=4.423) were between A (5 points) or B (4 points) on the Scoring Grid. The majority of these candidates (54.478%) scored ≥ 4.5 for Domain 1, whereas slightly less than half of them did (47.762%) for Domain 3. They displayed the ability to apply "broad" to "comprehensive" knowledge/concepts and to "interpret and analyse" (Scoring Grid) the interplay among different perspectives. At the other end of the scale, all Level 1 candidates scored 1 point (M=1.000), which is described as having "elementary knowledge" and being able to "identify simple relationships from a few perspectives". Therefore, a statistically significant decrease in scores from Levels 5 to

1 for these two domains indicated a drop in "complexity of thoughts", which was the scoring criterion stipulated on the Scoring Grid.

As for Domain 6 *Cultures/Values*, statistically significant differences in scores were found between all levels, except Levels 1 and 2 (Tables 5.10 and AII-1B-2, AII-1B-3, Appendix II). As shown on the Scoring Grid (Table 5.11), the performance in this domain was differentiated with reference to the range of cultures/values taken into consideration in formulating arguments. Similar to the differentiating criterion for Domains 1 and 3, the consideration of more views/ cultures/ values also involves a higher complexity, which is in line with the New Taxonomy. The decreasing scores for this domain from Levels 5 to 2 suggested the application of less complex skills in this aspect.

Table 5.11: The Scoring Grid of Domain 6 Cultures/Values

Domain of Skill	**Description of the mastery of the skill**				
Cultures/ Values	show appreciation of different cultures/ universal values; or shows empathy/ open-mindedness/ tolerance towards a wide range of people/ incidents/ views / values in the formulation of arguments	consider particular cultures/ universal values; or show empathy/ open-mindedness/ tolerance towards particular groups of people/ types of incidents/ views/ values in the formulation of arguments	show limited awareness of different cultures/ universal values, the concerns/ situations of different groups of people in the formulation of arguments	elaborate on their own views based on their own/ values/ cultures; without sound justification	
	6A	6B	6C	6D	

However, the quantitative analysis above just provided a possible explanation and a broad description for the differentiation in the performance by the scoring criteria for Domains 1, 3 and 6. The alignment between the taxonomies and actual performance at various Levels can only be established by examining the live scripts qualitatively for evidence of the complexity of *Knowledge* applied.

Besides, there is another question that has remained unanswered: can we conclude that the assessment of candidates on the *Knowledge* on *Cultures/Values* at Levels 1 and 2 did not conform to the cognitive model and substantive validity cannot be justified? To answer this question, instead of referring to the Domain of *Knowledge* in the New Taxonomy alone, the *Cognitive System* should also be examined. As postulated by Marzano et al. (2008), *Knowledge* is applied in synthesising arguments at Level 4 *Knowledge Utilisation* in the New Taxonomy. *Knowledge*, including the knowledge and concepts from different perspectives and of *Cultures/Values*, should be incorporated in the formulation of arguments. To align with the cognitive models, the examination should award higher levels to candidates who showed better performance in a higher-order skill, *Knowledge Utilisation* (which is equivalent to *Synthesis* in Bloom's Taxonomy according to Marzano et al. (2008)), whereas lower levels awarded to those who can *Analyse*, *Comprehend* or *Retrieve* information only. The differentiation of these cognitive skills by the examination will be analysed qualitatively in the following section.

Another observation for Domain 6 *Cultures/values* was that the mean score for Level 5 candidates was 3.717 (Table 5.10), closer to Grid Square 6B (3.6 points) rather than 6A (5 points). Only 13.386% of the candidates were awarded ≥4.5, closer to Grid Square 6A, which is described as being able to "show appreciation …towards a wide range of people/ incidents/ views/ values in the formulation of arguments" (Grid Square 6A on the Scoring Grid, Table 5.11). The lower scores for this domain might be explained by the optional application of this skill in the examination, as suggested in the content analysis in Chapter 4. Besides, the qualitative analysis in the following section for *Knowledge Utilisation* in the *Synthesis* of arguments by considering different *Cultures/values* may further shed light on the appropriateness of the demand for Domain 6 in the examination.

In short, the ANOVA analysis showed the differentiation of the performance in Domains 1, 3 and 6 by the examination between various Levels of Performance, with the exception of Domain 6 *Cultures/Values* in Levels 1 and 2. Nevertheless, the descriptors on the Scoring Grid merely provide some clues to the performance of candidates as reflected by the mean scores. Direct evidence for the alignment with the cognitive models from a thematic analysis of live scripts is also necessary for the investigation of the substantive validity of the examination, albeit a generalisation of the alignment between the examination and the taxonomy is not viable. The findings from the qualitative thematic analysis will be discussed in the following section.

5.3.1.2 Qualitative analysis

In the thematic analysis, scripts from five candidates [50] were analysed to examine the differentiation of the performance on *Knowledge*.

The performances became weaker from Candidate A to Candidate E (Level 5 to Level 1)[51], in terms of the complexity of *Knowledge,* which was put forward by Marzano et al. (2008) as a differentiating criterion for the Domain of *Knowledge* in the New Taxonomy. Candidate A (Level 5) demonstrated an ability to command the types of *Knowledge* at the top of the hierarchy, *Generalisations* and *Principles* (relationships), which were more prominent in his/her answer for Paper 1 Questions 2(a) and 3(b) (Table 4.3). S/he made use of relationships regarding a wide range of perspectives (including, the personal, social, economic, cultural and environmental aspects of

[50] *For Levels 2 and 1, even though the answer to each question was taken from different candidates, for an easier reference to them, they were named Candidates D and E respectively. These examples were typical examples of performance. Samples in the other 3 levels were also "typical" since they were awarded similar scores by the examiners. Being typical examples and analysed by question, the samples are comparable across levels.*

[51] *Further details of analysis are found in Tables AIII-1a to AIII-1e (Appendix III).*

issues in the local or global contexts) to elaborate on his/her arguments. In Paper 1 Questions 2(a), the candidate elaborated on the reasons for young people to undergo plastic surgery with reference to the *Generalisation* on the influence of individualism and freedom from western culture (L5.13)[52], as well as the changes in social norm brought about by the promotion of celebrities (L5.15), integrating knowledge of global culture and moral considerations.

> *L5.13: "under the flow of western culture of individualism and freedom, plastic surgery has become very common and acceptable in the society" (P1Q2a)*
>
> *L5.15: "…with the promotion of celebrities through mass media, the social norm change(s) to accept plastic surgery and believe that it is a way to boost self-esteem" (P1Q2a)*
>
> *(Candidate A, Table AIII-1a, Appendix III)*

Furthermore, in Paper 1 Question 3(b), s/he explained the global concerns arising from tourism by the "negative consequences" of global warming and the relationship between conflicts and "dissatisfaction towards tourists" (L5.23 and L5.24), drawing on knowledge of global environmental problems and social conflicts. Candidate A was therefore able to integrate *Generalisations* and *Principles* from various perspectives to conjure up an answer, showing more complex "mental procedures" in the terms of Marzano et al. (2008).

> *L5.23: "Since negative consequences led by global warming like extreme weather or rising sea level is threatening the whole world, different countries will have the concern on carbon emission resulted in (from) international tourism." (P1Q3b)*
>
> *L5.24: "These conflicts will lead to growing dissatisfaction towards tourists or even damaged the cultural or historical relics." (P1Q3b)*
>
> *(Candidate A, Table AIII-1a, Appendix III)*

Even though the application of higher-order *Knowledge*, *Generalisations* and *Principles*, was also found in the answers of Candidates B, C and D (who attained Levels 4, 3 and 2 respectively),

[52] *Corrections to spellings or grammatical errors were added in brackets.*

more limited perspectives were considered by them and coherence was lacking in the elaborations. Referring to the answers for Paper 1 Question 2(a) again for a comparison with the performance of Candidate A, Candidate B attempted to explain the reasons for young people undergoing plastic surgery from the perspective of personal development, suggesting that being teased by others leads to low self-esteem (L4.12). However, the linkage between low self-esteem and plastic surgery was omitted.

> *L4.12: "they will be teased or treated unequally in the society. Therefore, the(ir) self-esteem is low…" (P1Q2a)*
>
> *(Candidate B, Table AIII-1b, Appendix III)*

As for Candidates C and D, they did not articulate clearly the *Generalisation* about the influences of celebrity and the mass media on young people (L3.8 and L2.6). While Candidate C did not show any misunderstanding of the concepts used, Candidate D showed a partial understanding of the influence of the mass media on young people. Being easily influenced is a general characteristic of young people, rather than a characteristic caused by the mass media as Candidate D put it. S/he blurred the *Generalisation* even more by adding the phrase, "without knowing (the) truth behind these promotion", which did not focus on the reasons for undergoing plastic surgery.

> *L3.8: "…they think plastic surgery is acceptable as celebrities accept it." (P1Q2a)*
>
> *(Candidate C, Table AIII-1c, Appendix III)*

> *L2.6: "Due to social mass media the youngsters are easily influence(d), without knowing truth behind these promotions…" (P1Q2a)*
>
> *(Candidate D, Table AIII-1d, Appendix III)*

The lack of coherence was not only evident in the elaborations on higher-order *Knowledge*, but also in the description of *Facts* and the use of *Vocabulary Terms*. Candidates B, C, D and E showed attempts to use some examples to answer the questions. However, they merely described the facts in a detached manner, thus they failed to integrate them into the explanation of their answers. Due

to their inability to make use of *Knowledge* in their arguments, isolated *Vocabulary Terms* or definitions of some terms or concepts were found in the answers of Candidates B, C, D and E. For instance, Candidate E just explained briefly legislation and freedom of expression, without relating these concepts to the main topics of the questions (Paper 1 Question 2(b) and Paper 2 Question 1(a)) (Table 4.3), which were about a ban on plastic surgery and factors for freedom of expression respectively (L1.4 and L1.5).

> *L1.4: "legislation…will be strictly forbiddening (forbidding) the teenagers" (P1Q2b)*
>
> *L1.5: "The different media show different views and some time invite or interview some citizen(s) to tell their views on different things.… It is one of the way of citizens use their freedom of expression." (P2Q1a)*
>
> *(Candidate E, Table AIII-1d, Appendix III)*

These samples illustrated the performance as expected in the Level Descriptors, in which only Level 5 candidates were able to "interpret and analyse coherently from different perspectives". Coherence in the incorporation of *Knowledge* in the answers was also one of the scoring criteria adopted by the examiners, indicating the alignment in the scoring process with the Level Descriptors. In Nominal Group Discussion 3, the examiners concurred that the scripts from Level 4 downwards did not elaborate on ideas coherently. Examiners E2 and E3 described the general performance of both Levels 4 and 3 candidates in the joint study as having a lot of "gaps" and failing to provide elaborations in relation to the question (Table 4.3): (The incoherence shown in the script will be further discussed in Section 5.3.2 on *Synthesis*.)

> *E2: "…They did not logically elaborate on their ideas, leaving some gaps and undermining the flow of the answer."*
>
> *E3: "Too general… He mentioned (in Paper 1 Question 3(b)) that 'the rise in the number of tourists caused some damage'. Then he continued by referring to 'heaps of rubbish'. He jumped to 'a serious destruction to the environment'. But how was that related to the damage to civilisation? There are a lot of gaps. Further down… 'the concern on a clean image of the country'…What is the concern? Which country? He was unable to relate it to the sources."*
>
> *(Nominal Group Discussion 3)*

In addition to coherence, the level of understanding is also a criterion for assessing students' level of *Knowledge*. Candidates B, C, D and E showed a lower level of understanding of the relevant terms and concepts. They used some inappropriate terms: for example enhancing "international image" (L1.13) as a benefit from tourism; non-existent terms: "intergenerational family" (L4.18), "in a harmony perspective" (L3.24), "living problems" (L1.12), as well as some wrong facts/ facts without evidential support (L3.21, L1.11), inappropriate generalisations on the impact of the loss in investment in primary industry (L3.23) and relationships between press freedom and confidence of the public (L2.14). Candidates D and E mistook press freedom in the question (Paper 2 Question 1) (Table 4.3) as freedom of speech, thus missing the point of the question (L2.12, L1.5).

> *L1.5: "The different media show different views and some time invite or interview some citizen(s) to tell their views on different things.… It is one of the way of citizens use their freedom of expression." (P2Q1a)*
> *L1.11: "…plastic surgery…will be creating a lot of deaths on the teenagers." (P1Q2b)*
> *L1.12: "living problems…Rents in urban areas have been increasing" (P1Q1b)*
> *L1.13: "Tourism enhancing 'international image" (P1Q3a)*
> *(Candidate E, Table AIII-1e, Appendix III)*

> *L2.12: "People have the right to speak up and it (is called) calls press freedom." (P2Q1a)*
> *L2.14: "A high degree of press freedom consolidate(s) the confidence of public' (P2Q1b)*
> *(Candidate D, Table AIII-1d, Appendix III)*

> *L3.21: "Environmental problems also harm the mental health of residents living nearby" (P1Q1b)*
> *L3.23: "China's economy is shifting to secondary and tertiary industry, loss in foreign investment in primary industry is a bearable cost." (P1Q1c)*
> *L3.24: "Second, negative impacts is brought in a harmony perspective" (P1Q3b)*
> *(Candidate C, Table AIII-1c, Appendix III)*

> *L4.18: "This will cause the problem of intergenerational family (should be skipped generation family), as…the farmer will choose to leave the family to earn a living in urban. The kids and their parents will stay…" (P1Q1b)*
> *(Candidate B, Table AIII-1b, Appendix III)*

For Domain 6 *Cultures/Values*, the thematic analysis of the scripts (Tables AIII-2a to AIII-2e, Appendix III) showed *Generalisations* and *Principles of Knowledge* on values or cultures. The average score for Domain 6 for all questions Candidate A attempted was 4.019, a bit higher than the average for Level 5 candidates: 3.717 (Table 5.10). Candidate A was able to answer the questions with reference to a "wider range of cultures or values", as illustrated by the Scoring Grid, including *social cultures*: "rural-urban disparity" (L5.1), "cultural conflicts" (L5.2); *moral values*: "values toward beauty" (L5.3), "social norms" (L5.5); *economic values*: "free market" (L5.4); and *social values*: "press freedom" (L5.6), "social harmony" (L5.7), and "civil values and awareness" (L5.8). The skill involved in integrating high-order *Knowledge* on various values and cultures in the answer is as complex as applying knowledge from various perspectives. Therefore, these excerpts provided evidence for complex mental processes in the *Knowledge* Domain as postulated in the New Taxonomy (Marzano et al., 2008).

L5.1: "social disharmony arised by (caused by) urban-rural disparity and migrant workers" (P1Q1b)

L5.2: "the second concern is the cultural conflicts arised from (caused by) international tourism." (P1Q3b)

L5.3: "Secondly, in terms of addressing the root problem, passing law is not dealing with the root cause of teenagers undergoing plastic surgery which is gaining peer recognition and incorrect values toward beauty." (P1Q2b)

L5.4: "If the government intervene(s) (with) the free market by passing laws to bar business opportunities, (the) profit of these companies may drop and they may oppose to the government" (P1Q2b)

L5.5: "However, with the promotion of celebrities through mass media, the social norm change(s) to accept plastic surgery and believe that it is a way to boost self-esteem…" (P1Q2a)

L5.6: "Press freedom include(s) freedom of expressing ideas or reporting news, be it positive or negative" (P2Q1a)

L5.7: "…press freedom help(s) expressing (express) social discontent, improving social harmony" (P2Q1b)

L5.8: "(a) high degree of press freedom can play the role of educator or promotor, helping the government to inculcate correct civil values on citizens, raising civil awareness." (P2Q1b)

(Candidate A, Table AIII-2a, Appendix III)

In contrast to the performance of Candidate A, the variety of values the other four candidates referred to in their answers was narrower. Only personal values were considered by Candidates C, D and E, while Candidate B was able to incorporate social values other than personal ones. Besides being able to draw a reasonable *Generalisation* of the acceptance towards plastic surgery in society (L4.5), Candidate B showed an understanding of the protection of the legal rights of the media (L4.6). Nevertheless, the G*eneralisation* was not integrated well in the explanation of the factors affecting press freedom.

L4.5: "Plastic surgery is now generally well accepted surgery in society as it doesn't do harm to others, while having few benefits to the teenager(s) himself." (P1Q2b)

L4.6: "the right of the press is well protected by law. ... Therefore, even the mass media spread something negative to the government, the government cannot sue the newspaper as long as the newspaper doesn't violate (the) law." (P2Q1a)

(Candidate B, Table AIII-2b, Appendix III)

As for Candidates C, D and E, social, economic and political values were either not deployed or not deployed appropriately by them. Therefore, Paper 2 Question 1 (Table 4.3), which is by nature on socio-political issues, was not discussed in a comprehensive manner by these candidates. Candidates C and D did not clearly articulate the *Generalisations* and *Principles* in relation to *Cultures/Values*, similar to their performance of other dimensions of *Knowledge*. For instance, Candidate C made an attempt to generalise the characteristics of mental immaturity of under-18s, but s/he did not make clear how value development could be linked to decision-making and adversity-handling in relation to plastic surgery (L3.3). Similarly, Candidate D failed to explain clearly the *Generalisation* about decision-making and identity development of young people (L2.2). Candidate E showed inadequate understanding of moral values by deploying an inappropriate phrase "to facilitate their personal images" (L1.3) as a reason for undergoing plastic surgery.

L3.3: "Under-18s are not mentally mature to take the surgery as they are still developing their values and critical thinking ability, as the surgery is permanent…as they may make decition (decisions) too easily and is not mentally strong enough to deal with possible side effects and failure." (P1Q2b)

(Candidate C, Table AIII-2c, Appendix III)

L2.2: "However, they tend to mature and would like to make their own decisions, since they are still at that age, where they are identifying themselves and they are easily influence(d) by the mass media." (P1Q2b)

(Candidate D, Table AIII-2d, Appendix III)

L1.3: "…undergo the plastic surgery it is because that they want to boost their confidence in social gatherings and to facilitate their personal images in order to get in close with their friends and to rebuild their relationships" (P1Q2a)

(Candidate E, Table AIII-2e, Appendix III)

Furthermore, the consideration of different value positions is categorised under Domain 6 in the Scoring Grid. In this respect, all five candidates, except Candidate D, mentioned counter-arguments. However, only Candidate A was able to conjure up sound rebuttals (L5.9), demonstrating an ability to synthesise arguments, which will be discussed further in Section 5.3.2.2.

L5.9: "Someone may also argued that high degree of press freedom is actually hindering government because it leads to and encourages demonstrations or strikes, and are opposing the government, which will then worsen relationship between the government and Hong Kong people….
However, such negative or dark side of government being reported is actually helping to improve the policy and government performance. For example, when the mass media reveal some negative side of a policy, the government can then fix the problem. Actually, the quality of policy is more important than it is not criticised by the public in terms of governance" (P2Q1b)

(Candidate A, Table AIII-2a, Appendix III)

Apart from displaying a narrower range of cultures/values in comparison with Candidate A, Candidates B, C, D and E did not show much application of higher-order *Knowledge* in relation to *Cultures/Values*. This supported the observation from the ANOVA analysis that the scale of the scores for this domain was squeezed towards the lower end.

From the qualitative analysis of the live scripts, evidence for the fulfilment of the Assessment Objective *n* (shown below) was found.

> *"• (n) to demonstrate an understanding and appreciation of different cultures and universal values;"*

Although cultural/value consideration is omitted in the Level Descriptors (as discussed in Chapter 4), Candidates A to E incorporated knowledge or understanding of cultures/values in their answers to a certain extent. Therefore, to fully reflect the Assessment Objective and to describe candidates' performance comprehensively, the performance in cultural/value consideration should be specified in the Level Descriptors.

5.3.2 Cognitive Skills: Information-handling, Synthesis and Evaluation

To evaluate substantive validity, Pearson's Correlations, a One-way ANOVA (Tables AII-5-2, AII-5-3, Appendix II) and an analysis of constructed binary variables (the dichotomised scores between pairs of these skills) were conducted to find out whether the examination differentiated candidates' Levels of Performance in accordance with the order of skills (*Analysis* (*Information-handling* in this study), *Synthesis* and *Evaluation*) stipulated by cognitive models. Evidence will also be drawn from the thematic analysis of live scripts.

5.3.2.1 Quantitative analysis

In the first place, following the research approach of Kreitzer and Madaus (1994) (as cited in Anderson & Krathwhohl, 2001), correlations among the scores for Domains 2 *Information-handling* (which is termed *Analysis* in Bloom's Taxonomy (1956) and the Revised Taxonomy

(Anderson & Krathwhohl, 2001)), 4 *Synthesis* and 5 *Evaluation* were computed to examine the differences in cognitive demands of these three skills. Referring to Table 5.12, while the majority of the domains were highly correlated with each other (>0.7), the correlation between Domains 2 *Information-handling* and 5 *Evaluation* was merely moderate (0.628).

Table 5.12 Correlations between Domains 2, 4 and 5 and other domains

Domains	2. Information-handling		4. Synthesis		5. Evaluation	
	Pearson Correlation	Number	Pearson Correlation	Number	Pearson Correlation	Number
1	0.794	252	0.890	360	0.816	358
2	1	252	0.763	252	0.628	250
3	0.789	252	0.902	360	0.810	358
4	**0.763**	252	1	360	0.814	358
5	**0.628**	250	**0.814**	358	1	358
6	0.619	238	0.768	345	0.781	344
7	0.775	252	0.870	360	0.779	358
8	0.826	252	0.888	359	0.800	357

Both the correlations and scatterplots (Figures 5.13, 5.14 and 5.15) pointed to a wider discrepancy in the scores of Domains 2 *Information-handling* and 5 *Evaluation* than that between the other two pairs. In all these three scatterplots, the majority of the points lay below the diagonals, indicating the mean scores for Domain 2 *Information-handling* being the highest and Domain 5 *Evaluation* the lowest among the three skills. Besides, a higher correlation (Pearson Correlation Coefficient=0.814) between Domains 4 *Synthesis* and 5 *Evaluation* may be explained by the similarity in cognitive demands. This tied in with the findings of Kreitzer and Madaus (1994) (as cited in Anderson & Krathwohl, 2001), suggesting that the cognitive demands of *Evaluation* and *Synthesis* being closer and that between these two skills and *Information-handling* being further apart.

Figure 5.13: A Scatterplot of Scores for Information-handling and Synthesis

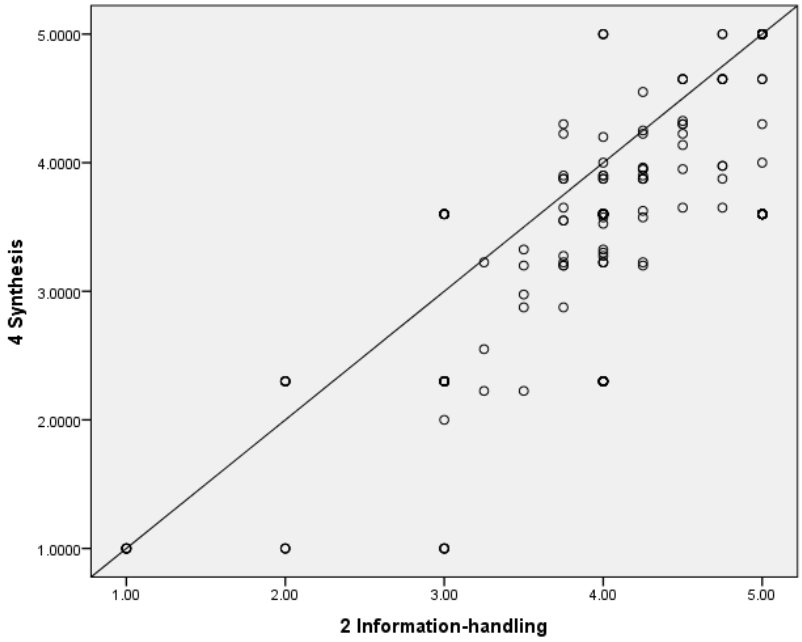


Figure 5.14: A Scatterplot of Scores for Information-handling and Evaluation

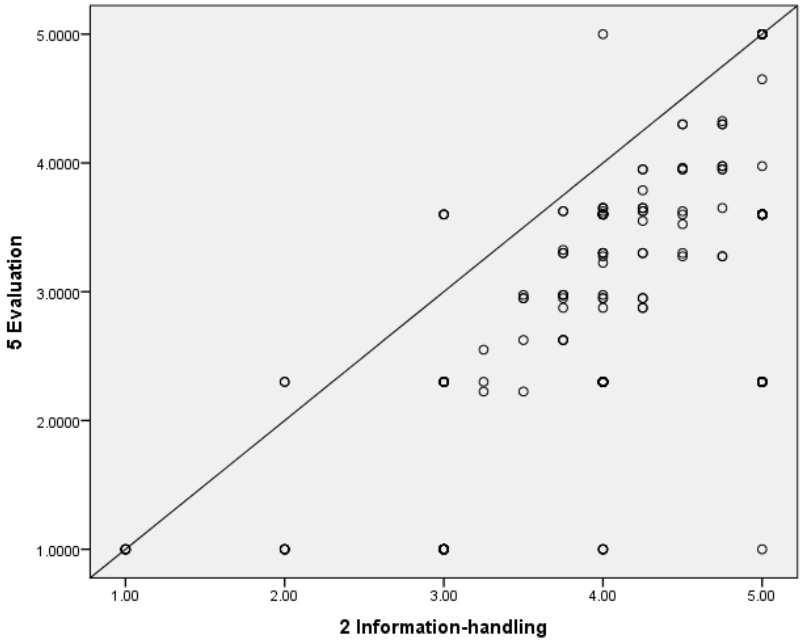
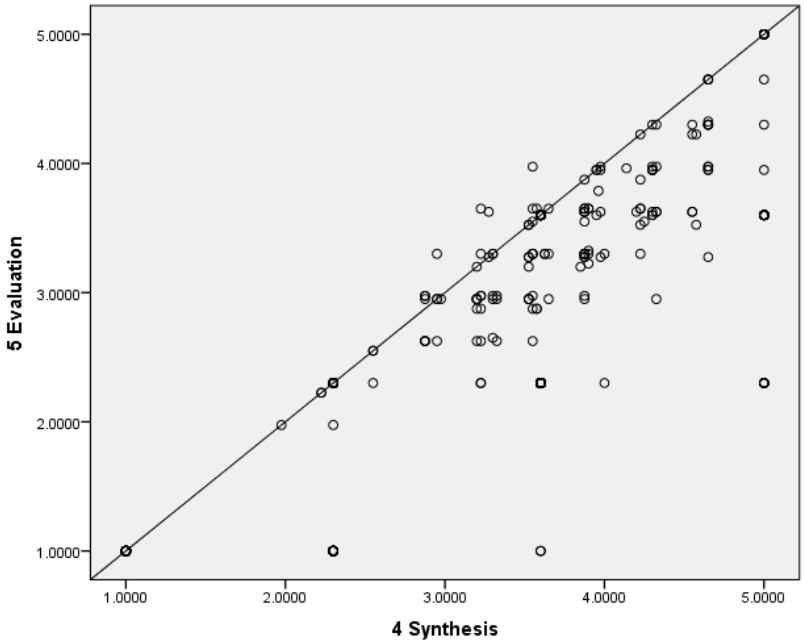


Figure 5.15: A Scatterplot of Scores for Synthesis and Evaluation



To further scrutinise the relative cognitive demands of Domains 2 *Information-handling*, 4 *Synthesis* and 5 *Evaluation*, binary variables were constructed by allotting 0 to a script not demonstrating the skill and 1 to a script demonstrating the skill. Assuming that the scores were awarded according to the Scoring Grid (Appendix I), Point 4 (B) or 5 (A) should be awarded for the demonstration of *Analysis*, Point 3.6 (B) or 5 (A) for *Synthesis* and Point 5 (A) for *Evaluation*. Therefore, in the construction of the binary variables, 1 was allotted to the average scores of *Analysis* >3, *Synthesis* ≥3 (about the mid-point between 2.3 (C) and 3.6 (B)), *Evaluation* ≥4.3 (mid-point between 3.6 (B) and 5 (A))[53]. The percentages of scores were tabulated in pairs of skills in Tables 5.16, 5.17 and 5.18.

From Table 5.16, the majority of scripts from Levels 4 and 5 showed both Domains 2 and 4, whereas less than half of the scripts from Level 3 did so ((2 *Information-handling*; 4 *Synthesis*) (1, 1): Level 3: 47.62%, Level 4: 79.76%, Level 5: 98.04%) (Table 5.16). As we move down from

[53] *The mid-points were taken as the cut off point for the construction of the binary variables as some examiners have awarded Bs to 4 Synthesis and As to 5 Evaluation, indicating the demonstration of the skills as perceived by the examiners.*

Level 5 to Level 3 a greater percentage of scripts showed the performance of Domain 2 but not Domain 4. While weaker candidates were observed to perform *Information-handling* rather than *Synthesis*, *Information-handling* is therefore perhaps less demanding than *Synthesis.*

From Tables 5.17 and 5.18, the majority of scripts at Levels 3 to 5 showed the mastery of Domains 2 and 4 only, but not Domains 5 ((2 *Information-handling*; 5 *Evaluation*) (1, 0): Level 3: 69.05%, Level 4: 88.10%, Level 5: 82.35%) (Table 5.17); ((4 *Synthesis*; 5 *Evaluation*) (1, 0): Level 3: 59.38%; Level 4: 81.08%; Level 5: 76.47%) (Table 5.18). It was shown that *Evaluation* was a skill too demanding even for the majority of candidates who attained Level 5. From the three tables, all these three skills, being higher-order as defined by Corliss & Linn (2011), were too demanding for Levels 1 and 2 candidates. Over 90% of them scored (0, 0) for all three pairs of skills.

The binary variables therefore give some indication of the relative cognitive demands of the three skills as stipulated in the Bloom's Taxonomy (*Information-handling*< *Synthesis* < *Evaluation*).

Table 5.16: The percentages of scripts showing Domains 2 Information-handling and 4 Synthesis for Levels 5 to 1 (0 indicating the skill not demonstrated; 1 for the skill demonstrated)

			2 Information-handling	
			0 (%(No.))	1 (%(No.))
Level 5	**4 Synthesis**	**0**	0.00 (0)	1.96 (2)
		1	0.00 (0)	98.04 (100)
Level 4	**4 Synthesis**	**0**	4.76 (4)	10.71 (9)
		1	4.76 (4)	79.76 (67)
Level 3	**4 Synthesis**	**0**	21.43 (9)	21.43(9)
		1	9.52 (4)	47.62 (20)
Level 2	**4 Synthesis**	**0**	91.67 (11)	8.33 (1)
		1	0.00 (0)	0.00 (0)
Level 1	**4 Synthesis**	**0**	100.00 (12)	0.00 (0)
		1	0.00 (0)	0.00 (0)

Table 5.17: The percentages of scripts showing Domains 2 Information-handling and 5 Evaluation for Levels 5 to 1 (0 indicating the skill not demonstrated; 1 for the skill demonstrated)

			2 Information-handling	
			0 (%(No.))	1 (%(No.))
Level 5	5 Evaluation	0	0.00 (0)	82.35 (84)
		1	0.00 (0)	17.65 (18)
Level 4	5 Evaluation	0	8.33 (7)	88.10 (74)
		1	1.19 (1)	2.38 (2)
Level 3	5 Evaluation	0	30.95 (13)	69.05 (29)
		1	0.00 (0)	0.00 (0)
Level 2	5 Evaluation	0	91.67 (11)	8.33 (1)
		1	0.00 (0)	0.00 (0)
Level 1	5 Evaluation	0	100.00 (12)	0.00 (0)
		1	0.00 (0)	0.00 (0)

Table 5.18: The percentages of scripts showing Domains 4 Synthesis and 5 Evaluation for Levels 5 to 1 (0 indicating the skill not demonstrated; 1 for the skill demonstrated)

			4 Synthesis	
			0 (%(No.))	1 (%(No.))
Level 5	5 Evaluation	0	2.21 (3)	76.47 (104)
		1	0.00 (0)	21.32 (29)
Level 4	5 Evaluation	0	16.96 (19)	81.36 (90)
		1	0.89 (1)	1.79 (2)
Level 3	5 Evaluation	0	40.63 (26)	59.38 (38)
		1	0.00 (0)	0.00 (0)
Level 2	5 Evaluation	0	100.00 (24)	0.00 (0)
		1	0.00 (0)	0.00 (0)
Level 1	5 Evaluation	0	100.00 (24)	0.00 (0)
		1	0.00 (0)	0.00 (0)

In fact, as illustrated by the Level Descriptors, Level 1 candidates were not expected to be able to formulate arguments or evaluate. Typical candidates at this level were expected to "list viewpoints" and "describe related information from their own viewpoints". *Synthesis* and *Evaluation*, being higher-order skills in the taxonomies, were too demanding for these candidates. Therefore, the performance of Levels 1 and 2 in the live script study was in line with the requirements of the examination as specified in the Level Descriptors.

Synthesis, being a skill judged of higher order than *Information-handling*, was performed by candidates attaining higher Levels of Performance. Based on the assumption that the scripts were

scored in accordance with the descriptions on the Scoring Grid derived from the Level Descriptors, the scores can be interpreted by the descriptions on the Grid. The description of Grid Square 4B on the Scoring Grid specified the ability to "synthesise their own opinions/ suggestions with partly reasonable arguments" (Appendix I). However, *Synthesising* was not expected at Grid Square 4C ("elaborate on opinions/ suggestions…" (Appendix I)). From Table AII-1B-3 (Appendix II), as interpreted according to the descriptions on the Scoring Grid, candidates at Level 3 or below were not able to synthesise as they scored below 4B (3.6 points) (Level 3: M=3.093; Level 2: M=2.300 (4C); Level 1: M=1.00) In other words, the results suggested the dividing line for mastering the skill of *Synthesis* lay between Levels 3 and 4. A higher-order thinking skill, *Synthesis*, being performed by candidates at Levels 4 and 5, provided evidence for the differentiation of cognitive performance of candidates by the examination in accordance with the cognitive taxonomies. However, the details of the performance of *Synthesis* cannot be evident from the scores. Therefore, the alignment of the assessment of *Synthesis* with cognitive models will be further analysed qualitatively in Section 5.3.2.2 with reference to the live scripts and the views of the nominal group.

In a nutshell, the statistics show that candidates' abilities to master thinking skills were differentiated by the examination in accordance with cognitive models. The correlation between *Synthesis* and *Evaluation* was higher than that between *Information-handling* and *Evaluation*, showing a closer relationship in the former pair of skills. The analysis of the constructed binary variables provided further evidence for an increase in demand from *Information-handling* to *Synthesis* and then *Evaluation*, as postulated in Bloom's Taxonomy. In other words, the examination distinguished candidates in accordance with the cognitive models, lending support to the substantive validity. Even though the scores were awarded with reference to descriptive criteria on the Scoring Grid, the variation in complexity of the cognitive skills applied by

candidates cannot be indicated by the scores. The details of the performance of the cognitive skills of various demands cannot be shown by merely the scores of the scripts, the thematic analysis in the following section will examine further the alignment of the examination with the cognitive models.

Another observation from the quantitative analysis was the discrepancy between the Level Descriptors and the performance of *Evaluation* at Level 5. The performance on *Evaluation* was the poorest (M=3.663, SD=0.704) (Table 5.9) among the three skills, pitching B only for the candidates at Level 5. In terms of the Scoring Grid (Grid Square 5B), most candidates compared "without clear criteria/standards". They were not able to "evaluate various viewpoints" as stipulated by the Level Descriptor for attaining Level 5. Although the Level Descriptors are designed to be deployed in a holistic manner, rather than a checklist of criteria for attainment, the inability of most candidates at the top level in the study to "evaluate … based on clear criteria" did indicate a discrepancy between the Level Descriptors and the actual performance of candidates, which is worth-noting for test developers. Is the demand for *Evaluation* set too high? This question will be addressed in the following section on qualitative analysis and the implications for test developers will be further discussed in Chapter 7.

5.3.2.2 Qualitative analysis

The statistical analysis provided evidence for the differentiation of performance on *Information-handling*, *Synthesis* and *Evaluation* between levels (except *Evaluation* between Levels 1 and 2) (Section 5.3.2.1) and higher average scores for *Information-handling* than *Synthesis* and *Evaluation*. However, the relative cognitive demands of these three skills cannot be fully reflected by an analysis of the scores with reference to the descriptions on the Scoring Grid. As such, a

qualitative analysis of the performance in authentic samples was conducted.

A higher-order *Information-handling* skill, *Generalising* trends shown in the data, was mastered by all the five candidates (Tables AIII-3a to AIII-3e, Appendix III). *Generalisation* was defined as "making a general statement" on the "patterns or connections" "from information that is already known or observed" (Marzano et al., 2008, p.19). According to Marzano et al. (2008), *Generalisation* is equivalent to Levels 4, 5 or 6 in the Revised Taxonomy and is a more demanding *Information-handling* skill than *Interpretation* or *Identification of information*. All the five candidates were able to point out the trends (i.e. the general changes in the percentage contribution of different types of industries to the GDP, the percentage of rural population and the incomes) for *Sannong* (agriculture, rural areas and farmers in China) in Question 1a (Table 4.3) and the trends shown in the tourist information (i.e. the general changes in the international tourist arrivals and tourism receipts in the world) in Question 3(a) (Table 4.3). Therefore, these samples provided clues for the relatively higher scores attained by candidates in *Information-handling* (Table 5.9). The difference in performance among the five candidates lay in the ability to generalise from sets of data in various presentation formats (tables of figures and a cartoon) and to use the data to describe the trends. Candidate A described the trends clearly in terms of the changes in the percentages calculated from the sources (in bold) for Question 1(a) (Table 4.3) (L5.1 and L5.3) and Question 3(a) (L5.4).

L5.1: "From Source A, contribution of primary industry, including farming, drop(ped) from 27.1% in 1990 to 10% in 2003, ***drop(ped) by 17.1%*** *in 23 years." (P1Q1a)*

L5.3: "Economic contribution, which is one tourism receipts also rise (rose) from 262 billion US dollars in 1990 to 1078 billion in 2012, the number has ***increased by almost 5 times*** *while that of tourist number has* ***rise(n) by more than 2 times.*** *" (P1Q1a)*

L5.4: "In Source A, (the) number of international tourist s arrived rise (rose) from 4.34 million in 1990 to 1035 million in 2012, showing a continuous rising trend in the 2 decades. Economic contribution… rised (rose)… the number has ***increased by almost 5 times*** *while that of tourist number has rise(n)* ***more than 2 times.*** *" (P1Q3a)*

(Candidate A, Table AIII-3a, Appendix III)

However, the other four candidates were unable to make full use of the data to describe the trends clearly for both Questions 1 and 3. The performance of Candidates B, C and D was quite similar in terms of *Generalisation.* They described the rate of increase or decrease by quoting the figures, without further calculations (L4.2, L3.1, L2.1 and L2.3). As for Candidate E, s/he was not able to make generalisations from all the data sets given in the question. Only one general trend was identified from Source A in Question 1(a) (L1.1). In Question 3(a), s/he merely quoted the figures in different years, without any details of changes, such as the rate or magnitude of change (L1.3).

*L4.2: "In the source, both tourist arrivals and tourism receipts **increase sharply** from 1990 to 2012. The arrivals increase from 434 million in 1990 to 1035 million in 2012, while the receipts increase from 262 million to 1078 million in (the) same period of time." (P1Q3a)*

(Candidate B, Table AIII-3b, Appendix III)

*L3.1: "... the percentage GDP decrease(s) drastically from 27.1% at 1990 to 10% in 2013, while that of secondary industry **increase(s) slightly** and that of tertiary industry **increase(s) largely**." (P1Q1a)*

(Candidate C, Table AIII-3c, Appendix III)

*L2.1: "According to Source A, the percentage contribution of primary industry were (was) **decreasing gradually** from 27.1% to 10.0% between 1990 and 2013, the percentage contribution of tertiary industry were (was) **increasing gradually** from 31.5% to 46.1%." (P1Q1a)*

(Candidate D, Table AIII-3d, Appendix III)

*L2.3: "The international tourist arrivals shown in Source A is **increasing sharply** from 434 million people in 1990 to 1035 million people in 2012." (P1Q3a)*

(Candidate D, Table AIII-3d, Appendix III)

L1.1: "With reference to Source A, there has been an overall decrease from 27.1% in 1990 to 10.0% in 2013, nearly by two thirds." (P1Q1a)

(Candidate E, Table AIII-3e, Appendix III)

L1.3: "For the international tourist arrivals, it has increased from 434 million in 1990 to 1035 million in 2012, while for the international tourism receipts, it increased from 262 billion dollars in 1990 to 1075 billion dollars in 2012." (P1Q3a)

(Candidate E, Table AIII-3e, Appendix III)

From the mean scores on the live scripts discussed in Section 5.3.2.1, *Synthesis* was suggested to be performed by candidates attaining Levels 4 and 5. The qualitative analysis provided further evidence to the performance of *Synthesis,* which is defined as "putting elements together to form a coherent argument" (Anderson & Krathwohl, 2001, p.31). *Synthesis* was identified in the answers of Candidates A and B (Tables AIII-4a to AIII-4e, Appendix III). Candidate A demonstrated the ability to synthesise coherent arguments in response to all questions. S/he considered counter-arguments and justified his/her stance by a strong rebuttal, fulfilling what constitutes a synthesis as suggested by Anderson & Krathwohl (2001). Firstly, in Paper 1 Question

2(b) (Table 4.3), s/he formulated a convincing rebuttal to the counter-argument that banning plastic surgery "can take effect in short time" (P1Q2L5.3 and L5.4).

> *P1Q2L5.3: "Someone may argue that the policy can take effect in short time to stop teenagers from undergoing plastic surgery, so as to prevent any medical accidents or negative consequence on the growth of teenagers in Hong Kong."*
>
> *P1Q2L5.4: "However, in (the) long run, it is not effective to change their values towards beauty and raise their awareness towards (the) danger of such (an) invasive procedure in surgeries. In fact, to solve such problem involving value judgement, soft measures should be used for long term effectiveness."*
>
> *(Paper 1 Question 2(b), Table AIII-4a, Appendix III)*

In Paper 2 Question 1(b) (Table 4.3), Candidate A again put forward a cogent argument with a sound rebuttal. S/he explained clearly how a high degree of press freedom, which allows reports on both the "positive and negative comments" on a proposed policy (P2Q1L5.1), will help "the government to formulate better policies and to reach consensus in the society easier" (P2Q1L5.3). S/he was also able to put forward a sound rebuttal (P2Q1L5.6 to L5.8) against the counter-argument that a high degree of press freedom "is actually hindering governance" (P2Q1L5.4).

P2Q1L5.1: "Secondly, press freedom can help reviewing government policies, improving the quality of government policies, improving the quality of government policies. For example, media will invite specialist(s) to express their view toward important policies like Third Runway or Solid Waste Charging in Hong Kong, including both positive and negative comments."

P2Q1L5.3: "Since the effectiveness of governance does not simply lies on the feasibility of government policies, but also whether these policies can reach or satisfy citizens' demand, and so high degree of press freedom also (with) social reflection on important issues, help the government to formulate better policies and to reach consensus in the society easier."

P2Q1L5.4: "Someone may argue that high degree of press freedom is actually hindering governance because it leads to and encouraged(s) demonstration or strikes, and are opposing the government, which will then worsen relationships between the government and Hong Kong People."

P2Q1L5.6: "However, such negative or dark side of government being reported is actually helping to improve the policy and government performance. For example, when the mass media reveal some negative side of a policy, the government can then fix the problem."

P2Q1L5.7: "Actually, the quality of policy is more important than it is not criticized by the public in terms of governance."

P2Q1L5.8: "For example, the policy of building 85000 housing flat(s) suggested by former Chief Executive Tung Chee-hwa, though is enforced and implemented, face serious criticize (criticism) and opposition after that, leading to the resignation of him. So the quality of policy is more important."
(Paper 2 Question 1(b), Table AIII-4a, Appendix III)

Though Candidate B was able to formulate some arguments in response to the questions, his/her arguments were not as coherent as those of Candidate A, thus not fully fulfilling the requirement for *Creating* (*Synthesis*) as suggested by Anderson & Krathwohl (2001). The incoherence can be illustrated by the following excerpts from the answer to Paper 2 Question 1 (Table 4.3). S/he made an argument of how a high degree of press freedom can reveal information of government's proposals and force the government to provide further explanation for the decision-making process (P2Q1L4.1 to L4.3). S/he tried to make use of an example to support his/her own argument. Nevertheless, there were some factual errors (instead of "getting permission to run his television programmes publicly" (P2Q1L4.2), Mr Wong applied for a free television broadcasting license).

S/he did not make clear how press freedom affects governance by "supervising the government" by using this example (P2Q1L4.1).

> *P2Q1L4.1: "First of all, high degree of press freedom help(s) supervise the government. When the government is doing anything or proposing measures, the mass media will keep checking of (on) it and spread it to the public. Therefore, when the government try (tries) to do something which hinders the interest of certain stakeholder, the mass media will disclose this and the public may respond to it."*
>
> *P2Q1L4.2: "For example, Wong Wai Kei is not allowed to get the permission to run his television programmes (broadcasting company) publicly in televisions due to a series of factors, in which the government didn't disclose because it's confidential. The mass media record it and spread it to the public. Therefore, the public think that the acts of government are not transparent enough."*
>
> *P2Q1L4.3: "This arouse(s) discontent of the public and people oppose the government, forcing her to give a detail explanation."*
>
> *(Paper 2 Question 1(b), Table AIII-4b, Appendix III)*

As discussed earlier, coherence in the argument was also emphasised by the examiners when scoring the live scripts. The answers from Level 4 candidates were described by the examiners as providing arguments with "a lot of gaps" (as quoted in the following). In a sample of Paper 1 Question 3(b) (Table 4.3) discussed in the meeting, Candidate F (Level 4) was able to "synthesise" an argument about the environmental problems brought by an increase in tourists. However, without logical linkages between some of the sentences, s/he failed to clarify why the environmental problems deserve global concern and how the problems s/he described will affect the image of countries. As a result, s/he failed to focus on the core of the question.

> *E2: "…They did not logically elaborate on their ideas, leaving some gaps and undermining the flow of the answer."*
>
> *(Nominal Group Discussion 3)*

Candidate F: Paper 1 Question 3(b) (translated from Chinese)
"Because of the rise in the number of international tourists, the number of people travelling increases greatly. If tourists are mostly uncivilized, the destruction to heritage or the environment may be hard to estimate. For example, if tourists litter, (or) because of the rise in the number of tourists, rubbish heaps may be formed, affecting the environment. For the cleanliness and the image of countries, undoubtedly, the civilised tourism is becoming the concern of the world."

Evaluation, being more demanding than *Synthesis* and *Information-handling* as postulated by Bloom (1956), was only demonstrated by Candidate A in his/her answers to all questions. The candidate assessed the effectiveness of the measure of "providing subsidy" to farmers on their living standard and social harmony in Paper 1 Question 1 (P1Q1L5.1). In Paper 1 Questions 2 and 3 (Table 4.3), s/he evaluated whether the ban on plastic surgery can solve the "root problem" in the long run (P1Q2L5.1 and L5.2) and the scale of the impact of global warming to justify it as a concern for the world (P1Q3L5.3).

P1Q1L5.1: "So providing subsidy can solve the problem of social disharmony as migrant workers drop and living standard of farmers rised(rose)."

P1Q2L5.1: "Passing laws can only lead to behavioural change but not attitude change. In (the) long run, when the youngster(s) reach 18 years old, they will still undergo plastic surgery."

P1Q2L5.2: "…in terms of addressing the root problem, passing law is not dealing with the root cause of teenagers undergoing plastic surgery which is gaining peer recognition and incorrect values toward beauty."

P1Q3L5.3: "Since negative consequences led by global warming like extreme weather or rising sea level is threatening the whole world, different countries will have the concern on carbon emission resulted in(from) international tourism."

(Table AIII-4a, Appendix III)

In Paper 2 Question 1, s/he evaluated the impact of a high degree of press freedom on governance with reference to clearly defined criteria for measuring governance: "implementation and enforcement of policy", "consensus in the society" in the "formulation of policies", "oppositions to the government", "credibility of the government". S/he justified his/her stance on the impact

by assessing the relative importance of the criteria for effective governance: the quality of policies and oppositions to the government (P2Q1L5.1 to L5.8, Table AIII-4a, Appendix III). S/he was able to make "judgements based on criteria", demonstrating the evaluation skill as defined by Anderson & Krathwohl (2001) (p.31).

However, from the quantitative analysis, most of the candidates in the study did not perform *Evaluation* as well as Candidate A and were not able to evaluate with reference to some clear criteria. Among candidates who have attained Level 5, the mean score for *Evaluation* was 3.667 only (Table 5.9), closer to Grid Square 5B (3.6 points). Was the bar raised too high for *Evaluation* in the examination? As this is the definition of the skill in the Revised Taxonomy (Anderson & Krathwohl, 2001), the requirements should not be lowered just because candidates cannot perform it. For a public examination to be valid, the alignment of the requirements with academically acceptable standards is of paramount importance. In fact, from the nominal group discussion, examiners unanimously concurred with this requirement for awarding 5 points to Domain 5.

> *E1: "…For evaluation, they need to put forth some criteria and then weigh the relative importance of different aspects. They were weak in putting forth some criteria for assessment."*
>
> *E3: "They cannot assess the relative importance of factors. For example, they may put forth several reasons, but they cannot explain which one is a more important reason."*
>
> *(Nominal Group Discussion 3)*

Candidate B's answer (Table AIII-4b, Appendix III) illustrated how a candidate may be able to *Synthesise*, but not *Evaluate*. As discussed above, in Paper 2 Question 1 (Table 4.3), s/he made some arguments about some positive impacts of a high degree of press freedom on the policy-making process (P2Q1L4.1 to P2Q1L4.3, Table AIII-4b, Appendix III), showing his/her ability to synthesise. However, as the criteria for determining the effectiveness of governance was not

clearly delineated, the impacts on governance were not weighed. An impact assessment is hardly convincing without weighing the positive and negative impacts. S/he merely explained briefly that media reports may "draw (the) attention of certain stakeholders" who "resist the decision of the government", thus hindering "the implementation of policies" (P2Q1L4.4). A sudden shift from the negative impact - posing "hindrance to policy implementation", to the positive impact - implementing an "adjustment of measures" "to meet the demands of different stakeholders" (P2Q1L4.5), did not provide grounds for the judgement of the overall impact of press freedom on the effectiveness of governance.

> *P2Q1L4.4: "Some say that a high press freedom may hinder the implementation of policies as the negative opinion spreaded (spread) will always draw attention of certain stakeholders to resist the decision of government."*
>
> *P2Q1L4.5: "However, this procedure is in fact help(s) the government to understand the opinions of different stakeholders. So the government can adjust their measures or explain publicly with reasons so as to meet the demands of different stakeholders in public."*
>
> *(Table AIII-4b, Appendix III)*

The ability of *Synthesising* was hardly found at all in the other candidates. Candidate C was not able to clearly articulate the arguments in response to the questions. In Paper 1 Question 3 (Table 4.3), s/he failed to point out the impact of "conflicts between nations" and "global harmony". It was too far-fetched to mention "anti-globalization movements" as a consequence of conflicts with tourists. The argument was further blurred by the sentence "This is bad to … lowering national hatred to each other".

> *P1Q3L3.2: "This intensify(ies) conflicts between nations or regions, may even cause anti-globalization movement or movement anti-tourists from certain country. This is bad to global harmony and lowering national hatred to each other."*
>
> *(Table AIII-4c, Appendix III)*

In Paper 2 Question 1 (Table 4.3), Candidate C was also unable to relate press freedom to governance, though s/he described some possible effects of a higher and a lower degree of press

freedom. As shown in Excerpts P2Q1L3.1 and L3.2, s/he attempted to explain the positive impact by an inappropriate description of the role of the press: "free press act(ing) as coordinators between (the) government and citizens". Therefore, s/he failed to discern the evaluation criterion. Similar to Candidate B, s/he tried to explain briefly a counter-argument (P2Q1L3.3), but the rebuttal was not understandable. Though s/he appeared to be aware of the need to put forward some evaluation criteria, such as whether the effects are "long lasting" (P2Q1L3.5 and P2Q1L3.6), s/he was unable to make meaningful judgement according to these criteria.

> *P2Q1L3.1: "Second, in terms of smoothly implemented, free press act as coordinators between government and citizens , explain the policy to citizens to make citizens have a better understanding about the policy, and can analysis benefit and cost themselves rationally."*
>
> *P2Q1L3.2: "With low press freedom, citizens lost trust about "facts" on press and media lost function as coordinator and cannot help smooth implementation of policies."*
>
> *P2Q1L3.3: "People may argue that press freedom let people know the dark side of government, cultivate anger and discontent toward government, hence make social movements happens (happen) more frequently and drag back governance efficiency."*
>
> *P2Q1L3.4: "This is true but it should be noticed that this happens only at the early stage desition (decision making) of governance."*
>
> *P2Q1L3.5: "The effect is short. Once (the) government amend(s) (the) policies, social movement would stop."*
>
> *P2Q1L3.6: "But with low freedom of press, the negative effects on governance efficiency can be long lasting, such as making ineffective policies, having government official with low working abilities."*
>
> *(Table AIII-4c, Appendix III)*

Both Candidates D and E performed much worse. Among these two, Candidate D performed a bit better by making an effort to explain briefly in response to the questions, whereas Candidate E mainly described the data given in the questions and failed to elaborate on arguments with a clear focus on the gist of the questions. Their performance can be illustrated by the answers to Paper 2

Question 1(b). Both mistook freedom of expression as equivalent to press freedom, thus the arguments were not focused on the impact of press freedom on governance. Candidate D showed effort in responding to the question directly by explaining briefly the effects of freedom of expression on the smooth implementation of policies (P2Q1L2.2 to L2.4). However, Candidate E just repeated some phrases from the information given in the question, for example, "less corruption, more efficient administration" (P2Q1L1.1) and explained freedom of expression briefly, failing to formulate grounded arguments in response to the question. Therefore, the performance fitted into the description of Grid Square 4D on the Scoring Grid.

> *P2Q1L2.2: "Although, HKSAR allows public discussion during the consultation stage but most public cannot ensure their voice are actually being heard."*
>
> *P2Q1L2.3: "Thus, the high degree of press freedom ensure(s) public speak out their opinions towards policy and make adjustment to the proposals."*
>
> *P2Q1L2.4: "As public's voice (has) been heard, the policy carried out will run more smoothly and effective and cause less problem."*
>
> *(Table AIII-4d, Appendix III)*
>
> *P2Q1L1.1: "In Source A, it say the more press freedom a society has, less corruption, more efficient administration, higher political stability…"*
>
> *(Table AIII-4e, Appendix III)*

Both Candidates D and E were awarded the lowest score for Domain 5. They provided a one-sided argument for the ban on plastic surgery in Paper 1 Question 2 (Table 4.3) (Tables AIII-4d, AIII-4e, Appendix III). In Paper 2 Question 1, Candidate D did not consider any counter-arguments, as described in 5D on the Scoring Grid, nor mentioning any negative impacts of a high degree of press freedom. Even though Candidate E pointed out some positive impacts (P2Q1L1.1 and P2Q1L1.2) and one negative impact (P2Q1L1.5), without a clear delineation of the impact of press freedom on governance, *Evaluation* skill was not shown.

P2Q1L1.1: "In Source A, it say the more press freedom a society has, less corruption, more efficient administration, higher political stability…"

P2Q1L1.2: "The higher degree of press freedom can benefit the economic, social and political. All of the benefit can increase the credibility of the government."

P2Q1L1.5: "Some of the people may say their view will make a lot of argue(ments) when there is a high degree of press freedom."

(Table AIII-4e, Appendix III)

In fact, the consideration of counter-arguments and formulation of rebuttals also reflects the performance of *Synthesis*. The one-sided arguments provided by Candidates D and E not only indicated their weaker *Evaluation* skill but also their *Synthesis* skill. The close linkage between these two skills is manifested in the correlation coefficients (Table 5.12) and the Scatterplots (Figure 5.15) in the previous section.

As shown in Figures 5.14 and 5.17, candidates scored much poorer on *Evaluation* in comparison with *Information-handling*. Candidate G (Level 2) scored 1 for *Evaluation*, but 4 for *Information-handling* in Paper 1 Question 3. Similar performance on *Evaluation* was found in the answers of Candidate E (Level 1). In Question 3(b), Candidate G merely pointed out the "local pollution problems" with the increase in the number of tourists. However, no effort could be traced for the assessment of the scale of the problem to justify it as a global concern. Perhaps the confusion of carbon dioxide as a local pollutant, rather than a greenhouse gas made it difficult for him/her to further examine the global impact.

Despite the inability in *Evaluation*, s/he was able to generalise trends from the data in Question 3(a): the magnitude of increase in the international tourist arrivals from 1990 to 2012. In Question 3(b) (Table 4.3), even though s/he misinterpreted the compound bar graph on carbon dioxide emissions in 2005 and 2035, s/he pointed out the increasing trend in the emissions: "The increase

in 30 years will be more than 313%". This sample provided a clue for a wide discrepancy in the cognitive demands between *Information-handling* and *Evaluation*, which is in line with the hierarchical ordering of the taxonomies. *Information-handling*, being less demanding, therefore could be better managed by candidates attaining lower Levels of Performance.

> ***Candidate G: Paper 1 Question 3(a) (translated from Chinese)***
> *"According to Source A, the international tourists arrivals showed a continuously increasing trend. It increased by more than 1 time from 0.434billion in 1990 to 1.035billion in 2012."*
>
> ***Candidate G: Paper 1 Question 3(b) (translated from Chinese)***
> *"Firstly, on the environmental aspect, the great increase in the arrivals of international tourists led to an unexpected local pollution problem. According to Source B, the carbon dioxide emitted from the tourist industry in 2005, especially from vehicles and other means of transport, was higher than the expected emissions in 2035. The expected emissions in 2035 will be 16%, whereas the emissions in 2005 was 29%. The increase in 30 years will be more than 313%. This showed that the great increase in the number of arrivals of international tourists will worsen the environmental pollution problem, arousing global concern."*

From the findings of the qualitative analysis on the performance of *Information-handling*, *Synthesis* and *Evaluation*, the candidates attaining Levels 4 and 5 (Candidates B and A) were able to master higher-order *Information-handling* skills and *Synthesis*. As for *Evaluation*, which was suggested to be the most demanding by the quantitative analysis, only Candidate A was able to fulfil the requirements for this skill. This is in line with the relative cognitive demands of these three skills as postulated by Bloom (1956).

5.4 Chapter Summary

In Chapter 5, both quantitative and qualitative evidence was solicited in support of Validity Argument *(2) The Level Descriptors appropriately differentiate the performance of candidates.* The appropriateness was determined by the differentiation of performance by the Level Descriptors and the alignment with cognitive models.

Firstly, the ANOVA analysis showed statistically significant differences in the scores between students demonstrating different Levels of Performance for each of the skill domains (with the differences between Levels 1 and 2 of Domains 5 *Evaluation* and 6 *Cultures/Values* as the exceptions) and for their overall performance (average of all the 8 domains) of scripts. Since scores were awarded in accordance with the Scoring Grid derived from the Level Descriptors, evidence for an appropriate differentiation of the performance of candidates by the Level Descriptors was gathered for the justification of the substantive validity of the examination.

Following the ideas of Newton et al. (2014) and Pellegrino et al. (2001) to evaluate the validity of the examination with reference to the alignment with cognitive models, evidence has also been gathered for the differentiation of the performance of candidates by the complexity of cognitive skills. The correlation and analysis of constructed binary variables indicated that the cognitive demands of *Synthesis* and *Evaluation* were closer than that between these two skills and *Information-handling*. Candidates attaining higher Levels of Performance (Levels 4 and 5) were able to demonstrate higher-order skills in Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001) and the New Taxonomy (Marzano et al., 2008): *Synthesis* and *Evaluation*. The lowest-order skill, *Information-handling* among these three was observed to be performed by candidates attaining even Level 3.

The thematic analysis also provided evidence for the mastery of more demanding thinking skills by candidates attaining higher Levels of Performance. The qualitative analysis of the five samples indicated that the complexity of *Knowledge* (Domains 1, 3 and 6) decreased from Candidate A (Level 5) to Candidate E (Level 1), which was in line with the quantitative evidence.

Overall Candidate A (Level 5) was able to make use of more complicated *Knowledge*: *Principles* and *Generalisations* from a wider perspective, including personal, local, national and global perspectives, and in relation to a variety of social or economic cultures and values. Arguments were synthesised coherently by using high-level *Knowledge* and *information-handling* skills (i.e. generalisation of trends), justified by sound rebuttals. From both the quantitative analysis of the scores and the thematic analysis, only Level 5 candidates demonstrated the *Evaluation* skill, deploying clearly discerned evaluative criteria.

Candidate B (Level 4) was also able to *Synthesise*, though not in such a coherent manner as Candidate A did. The variety of perspectives and cultures/values taken into account in the answers was less than that of Candidate A. Further down the Levels of Performance, all the other three candidates in the thematic analysis did not articulate their arguments clearly and coherently in response to the questions. Their answers were one-sided, considering predominantly personal values, omitting local and global socio-political cultures or values. As for the answers of Candidates D & E (Levels 2 and 1), the use of low-level *Knowledge*, which might be erroneous, was evident.

The quantitative analysis went in line with that from the qualitative analysis, which illustrated how candidates may command *Information-handling* better than *Synthesis* and *Evaluation*. All the five candidates were able to make generalisations from the data in the questions, showing their

command of a high-level *Information-handling* skill. On the other hand, only Candidate A was able to *Synthesise* coherently and to *Evaluate*. As candidates who performed higher-order thinking skills (*Synthesis* and *Evaluation*) were awarded higher Levels of Performance in the examination, the differentiation of the performance aligned with the taxonomies of cognitive skills, providing evidence in support for the substantive validity of the examination and Validity Argument (2).

Nevertheless, the scores for Domains 5 *Evaluation* and 6 *Cultures/Values* did not show any statistically significant differences between Levels 1 and 2 candidates. The mean scores for these two domains were lower at all levels in comparison with that of other domains. Even scripts which attained Level 5 overall were awarded far below 5 points for these two skills.

The evaluation of the substantive validity in this chapter provided hints for the enhancement of test development, as well as teaching and learning. Although the key scoring criterion of *Evaluation* was justified by the definition postulated by Anderson & Krathwohl (2001) and the agreement among the examiners, the incapability of candidates to attain high Levels of Performance to command this skill should be addressed. As for the performance on *Cultures/Values*, despite the lack of requirements for *Knowledge* on a variety of cultures and values in some of the questions in the 2015 LS Examination, all candidates in the thematic analysis deployed a certain level of *Knowledge* in this aspect in their answers. Therefore, to better reflect the performance of candidates in this domain and the Assessment Objective, the understanding and application of "a variety of cultures and values" should be embodied in the Level Descriptors. The implications of this assessment validity evaluation process will be discussed further in Chapter 7.

After investigating into the differentiation of candidates' performance in this Chapter, the

substantive validity will be evaluated from another aspect: the assessment of higher-order thinking skills, in Chapter 6.

CHAPTER 6 ASSESSMENT OF HIGHER-ORDER THINKING SKILLS

This chapter aims to investigate Validity Argument *(3) The 2015 LS Examination assesses the higher-order thinking skills of candidates specified in the Level Descriptors.* As higher-order thinking skills are the key Assessment Objectives of LS, evidence of performance of these skills in the examination is necessary in the evaluation of substantive validity. In this chapter, an in-depth thematic analysis of the alignment of the sequential cognitive processes with Kuhn's model (Kuhn, 2001, 2005) was conducted based on the live scripts and the think-aloud protocols. To examine whether candidates adopted higher-order thinking processes, think-aloud protocols and the samples of live scripts discussed in the previous chapter will be analysed in terms of *Meta-level Knowing*, *Dispositions* and two cognitive strategies in the KPI model (Kuhn, 2001, 2005): *Analysis* (equivalent to *Information-handling* in the live script study) and *Argument (Making Inference* will be discussed under *Argument Formulation* because the levels of performance of both are suggested by Kuhn (2001, 2005) to be differentiated on the same criteria: the integration of evidence). Nevertheless, *Enquiry*, which is one of the strategies in the model, will not be discussed because the task of answering examination questions does not constitute a complete enquiry process. The initial stage of an *Enquiry*, which comprises the formulation of enquiry questions and the design of data collection, was not performed in the written examination.

In the think-aloud study, candidates were asked to work on Paper 1 Question 3(b), which requires them to make use of Source A: a graph showing the changes in the number of international tourists and the revenue from tourism; Source B: a bar chart showing the carbon dioxide emissions in 2005 and 2035; and Source C: a passage adapted from a newspaper report on the guidebook for travelling overseas issued in mainland China.

> ***Paper 1 Question 3(b):***
> *"With reference to the sources provided, identify and explain two global concerns arising from the trends in international tourism you described in (a)."*

Adopting the classification suggested by Corliss & Linn (2011), higher-order thinking skills are defined as "applying knowledge and procedures to solve complex problems" (p.221), including *Analysis* (will be referred to as *Information-handling* in the following to align with the Scoring Grid), as well as formulation of arguments by *Synthesising* and *Evaluating*. Therefore, these skills are examined in this section for the justification of the Validity Argument (3).

6.1 Information-handling

Firstly, for *Information-handling*, as discussed in the previous chapter, both the quantitative and qualitative data showed that candidates were able to master *Information-handling* better than *Synthesis* and *Evaluation*, which are of a higher order (Tables 5.16, 5.17, 5.18, AII5-2 and AII-5-3, Appendix II). In the think-aloud protocols, all the 10 participants, including the one attaining Level 3 indicated the ability to handle more complex *Information-handling* skills. All of them started off their answers by making generalisations from the data set, showing the mastery of a more complex *Information-handling* skill, as postulated by Marzano et al. (2008). Case 3 (Level 4) and Case 4 (Level 5) generalised the increasing trend of carbon dioxide emissions from Source B and related it to the increase in tourists shown in Source A.

Case 3:

1	*describe the trends in Source A*
2	*obviously an increase*
3	*1990 was 434*
4	*2012 was 1035*
5	*with a gradual change*
6	*down there, the trend of the profits matches with the one above*
	…
14	*obviously source B shows carbon dioxide emission*
15	*obviously related to environmental issues*
16	*obviously one of the global concerns is global warming*
17	*as it affects people in the whole world*
18	*we can see the increase*
19	*projecting to 2035*
20	*the increase was very big*
21	*and we can see that a great proportion of the carbon dioxide emission is related to transport*
22	*related to air transport*
23	*accommodation is related to tourism*

(Table AIV-3, Appendix IV)

Case 4:

3	*directly related to the graph*
4	*the major item in this graph is air transport*
5	*the predicted in 2035 increases greatly from 05*
6	*car transport drops a lot*
7	*the underlying meaning is that car transport represents short distance travel & air transport is for long distance travel*
8	*this represents as time goes by*
9	*the average travelling distance of tourists is very long*
10	*the problem brought about is air transport emits a lot of CO2*
11	*aggravating global warming*
12	*as one of the global concerns*

(Table AIV-4, Appendix IV)

Even a Level 3 candidate (Case 2) was able to identify the general trend of the tourist arrivals. This is in line with the findings in Section 5.3.2 that the higher-level *Information-handling* skill could also be commanded by candidates attaining a lower Levels of Performance. However, the examination differentiated *Information-handling* skills by the comprehensiveness in the data analysis as stipulated by the Level Descriptors. In comparison with cases attaining higher levels, Case 2 failed to analyse the data provided in the question comprehensively. The carbon dioxide

emissions data were not analysed, therefore failing to relate the data to a global level of problems. Instead of identifying the problem of global warming, s/he incorrectly suggested air pollution as the problem, drawing a wrong conclusion from the data analysis (Line 14).

Case 2:

2	*the trend was the increase in arrivals by twofolds*
3	*therefore…*
4	*when there are more people*
	…
14	*because there are more people, there is air pollution*
15	*because there is CO_2 emission in Source B*

(Table AIV-2, Appendix IV)

6.2 Argument Formulation

Subsequent to *Information-handling*, candidates deployed the findings from their data analysis in formulating arguments. The strategy of *Arguments* in the KPI model (Kuhn, 2001, 2005) is equivalent to the skills of *Synthesis*, *Evaluation* (Bloom, 1956 and Anderson & Krathwohl, 2001) and *Knowledge Utilisation* (Marzano et al., 2008). In contrast to the taxonomies which compare the complexity of different thinking skills, Kuhn (2001, 2005) ordered different levels of command of the Strategy of *Arguments* in terms of the use of evidence in justifications. The highest level of performance was defined as the ability to integrate evidence in the formulation of arguments. He also suggested that *Meta-level thinking* and *Dispositions*, including knowledge, concepts and values are involved in the formulation of *Arguments*. The *Arguments* shown in the think-aloud protocols and live scripts will be analysed with reference to the use of evidence in this section and the use of *Meta-level thinking* and *Dispositions* in Sections 6.3 and 6.4 respectively.

In Chapter 5, it was shown that candidates attaining Levels 4 or above were able to perform a higher-order thinking skill, Domain 4 *Synthesis*. From the quantitative analysis of the live script

study, Level 5 candidates demonstrated the ability to "conceptualise evidence or use sufficient examples" as described on the Scoring Grid (Appendix I) for Domain 7 *Evidence* (Mean=4.257, with a maximum of 5 on the Scoring Grid on a 4-point scale) (Table 5.2). The think-aloud study provided further evidence for the ability to *Synthesise* among candidates attaining high Levels of Performance. In the think-aloud study, all Level 5 or above candidates (Cases 4, 6, 7 and 8) were able to integrate evidence in their synthesis of *Arguments* (Tables AIV-4, AIV-6, AIV-7 and AIV-8, Appendix IV), fulfilling the requirement for the highest-level performance of this strategy as postulated by Kuhn (2001, 2005).

Cases 4 and 7 (Level 5 and 5** respectively) made use of some examples of tourist behaviours provided in the sources of the question, such as "peeing in the Golden Bauhinia" (Lines 43, Case 4) and "the carving of names in an ancient temple in Egypt as an example" (Lines 64, Case 7) to infer and explain the global concern on "greater cultural conflicts" (Lines 39, 48, 49 and 53, Case 4) and the "destruction of tourist spots in other countries" (Lines 69, 70, Case 7).

Case 4

39	*to develop tourism further even with greater cultural conflicts*
	…
43	*peeing in the Golden Bauhinia Square*
	…
48	*in Hong Kong, people may think that this should be done in the toilet*
49	*it is not appropriate to do it in public places*
	…
53	*Hong Kong people will be discontented with the mainlanders*

(Table AIV-4, Appendix IV)

Case 7

64	*using the carving of names in an ancient temple in Egypt as an example*
	…
69	*this concern may lead to a problem of destruction of tourist spots in other countries*
70	*the others cannot enjoy them*

(Table AIV-7, Appendix IV)

Whilst Case 8 (Level 5*), based on the analysis of the data on the surging number of tourists and

the carbon dioxide emissions, inferred and elaborated on the global problems of greenhouse effect (Line 60), including “a rise in sea level” (Line 68) and justified it as a global concern. Though the emission figures were not quoted, s/he related the general trend of emissions to the problems, integrating evidence in the argument on the global concern over greenhouse effects.

Case 8

58	*more pollution*
59	*and more emissions*
60	*causing the greenhouse effect*
	…
63	*we need to explain why this is global*
64	*carbon dioxide can be a local pollutant*
65	*that is air pollution*
66	*need to talk about global concerns by referring to the effect in the world*
67	*e.g. not only in local*
68	*but in the world, e.g. rise in sea level and so on*
69	*people in the world will be affected*
70	*as so this is a global concern*

(Table AIV-8, Appendix IV)

Despite the sketchy nature of the protocol of Case 6 (Level 5*), s/he indicated the intention of using some examples from the sources and his/her own examples to “explain how these conflicts arouse concerns” (Lines 47 to 51). In the post-interviews, both Case 6 (Lines 53 to 54) and Case 7 (Lines 143 to 149) pointed out that they used examples to make their arguments grounded and more convincing, which was in fact the criterion set forth by Kuhn (2001, 2005) for a high-level and more effective performance in reasoning.

Case 6

47	*there will be conflicts between tourists and the locals*
48	*quoting some examples from Source C*
49	*and then add some examples of my own*
50	*and then explain how these conflicts arouse concerns*
51	*affecting social harmony*
	…
53	*if there are no examples, there is no ground*
54	*it says "with reference to" and it's given to you for your use*
55	*This is my way of doing it*

(Table AIV-6, Appendix IV)

Case 7

	(Getting back to the answer, why did you emphasise the use of examples and evidence?)
143	*On one hand, schools emphasised evidence*
144	*secondly, if there is no evidence, it seems ungrounded*
145	*you don't have concrete evidence for support…*
146	*maybe, just assumptions*
147	*not convincing*
148	*so there must be examples*
149	*examples help to make your elaborations reasonable*

(Table AIV-7, Appendix IV)

Think-aloud protocols provided some insight into the thinking processes of some high-level performance in argument formulation. However, since the protocols are truncated in nature, the integration of evidence in the arguments could be more clearly manifested in the live scripts. Therefore, the live scripts for Paper 1 Question 3(b) will be examined with reference to the use of evidence in what follows.

Referring to Sample A (Level 5), this candidate demonstrated how to integrate evidence into his/her arguments. S/he analysed the figures of carbon dioxide emissions provided in the question and used the figures to illustrate the "double increase" (L5.1) in emissions. The data illustrating that the aggravation of the problem of global warming led logically to the conclusion of an "immediate concern on how to reduce CO_2 emission without sacrificing international tourism." (L5.1).

> ***Sample A: (Candidate A)(Paper 1 Question 3(b))***
> *L5.1: "With globalization, exchange and international tourism is expected to be more common and popular, and it is extremely common for international tourists for travel through (by) airplanes. Since negative consequences led by global warming like extreme weather or rising sea level is threatening the whole world, different countries will have the concern on carbon emission resulted in international tourism. Also, as shown in Source B, CO_2 emission will rise from about 1200 million tonnes in 2005 to more than 3000 in 2035, almost a double increase. So this will be an immediate concern on how to reduce CO_2 emission without sacrificing international tourism."*

In Excerpts L5.2 to L5.4, Candidate A illustrated the cultural differences by examples of tourists' behaviours in Egypt and in Hong Kong (from Source C) and an example of Thailand's customs (from his/her own knowledge). These examples formed the basis for the argument that cultural differences may lead to "growing dissatisfaction towards tourists" (L5.2) and give rise to conflicts, well-justifying the cultural conflicts as "a great concern for governments all over the world" (L5.4).

> ***Sample A: (Candidate A)(Paper 1 Question 3(b))***
> *L5.2: "In (the) cultural aspect, the second concern is the cultural conflicts arised from (caused by) international tourism. From Source C, there are mainland tourists carving on ancient temple in Egypt and urinate(urinating) into a plastic bag in Hong Kong. In fact, these behaviour(s) may be accepted by mainlanders themselves, but not to Egyptians and Hong Kongers. Thus, conflicts arised (arose) due to cultural differences. These conflicts will lead to growing dissatisfaction towards tourists or even damaged the cultural or historical relics. Under international tourism, which is becoming more popular as shown in Source A, no countries can prevent the inflow of foreign culture, and cultural conflicts is (are) inevitable since different nations must have different cultures."*
>
> *L5.3: "For instance, people in Thailand believe that touching people's head is impolite while people in the West believe it is a friendly behavior."*
>
> *L5.4: "So under these cultural conflicts, it will be a great concern for governments all over the world to try to reduce locals' dissatisfaction, educating tourists in adapting to local culture without sacrificing the economic benefit generated by international tourism."*

Candidate A also integrated evidence in assessing how the concerns are global in nature. S/he demonstrated the skill of using the data from the sources and examples (both from the sources and his/her own knowledge) to assess the urgency and scope of the impact of global warming and cultural conflicts in the justification of the global concerns. As shown in the following excerpt, s/he used the data of carbon dioxide emissions and examples of tourists' behaviours to explain the argument that global warming is an "immediate concern" (L5.1).

Sample A: (Candidate A)(Paper 1 Question 3(b))
L5.1: ... "Also, as shown in Source B, CO_2 emission will rise from about 1200 million tonnes in 2005 to more than 3000 in 2035, almost a double increase. So this will be an immediate concern on how to reduce CO_2 emission without sacrificing international tourism."

The skill of integrating evidence can be better illustrated by a comparison with the next lower level of synthesis of evidence, "simple corroboration" as termed by Kuhn (2001). From the protocol of Case 3 (Level 4) (Lines 16 to 27), the candidate merely described some data from the sources and pointed out the increase in carbon dioxide due to international tourism, without linking up the data with the global concern of global warming to formulate a coherent argument. In Kuhn's (2001) terms, s/he merely made a simple corroboration of the data in the answer. In contrast, Case 8 (Level 5*) integrated the rising trend in carbon dioxide emission in the explanation of why global warming deserves global concern (Lines 68 and 69).

Case 3

16	*obviously one of the global concerns is global warming*
17	*as it affects people in the whole world*
18	*we can see the increase*
19	*projecting to 2035*
20	*the increase was very big*
21	*and we can see that a great proportion of the carbon dioxide emission is related to transport*
22	*related to air transport*
23	*accommodation is related to tourism*
24	*tourists need to take airplanes*
25	*especially international tourists*
26	*transport is the major reason*
27	*to explain the global concern arising from the increasing trend in international tourism*

(Table AIV-3, Appendix IV)

Case 8

35	*carbon dioxide emissions.. Expected to rise*
36	*as usual…calculate it*
	…
41	*guess that air transport means more transnational (travel)*
	…
45	*there is an increase*
46	*that's all from air transport*
47	*air transport… aeroplanes is the most environmentally unfriendly means of transport*
48	*I know that it is about more and more people visited foreign countries…I think…*
	…
59	*and more emissions*
60	*causing the greenhouse effect*
	…
68	*but in the world, e.g. rise in sea level and so on*
69	*people in the world will be affected*
70	*as so this is a global concern*

(Table AIV-8, Appendix IV)

Cases 4, 6, 7 and 8 and the answer of Candidate A indicated how candidates attaining Levels 5 or above were able to deploy high-level skills of *Argument* as suggested by Kuhn (2001 and 2005) by integrating evidence either from the *Analysis* of the data provided or their own experiences. However, the Level 4 candidate (Case 3) just indicated a simple corroboration of the sources, which is a lower level of the use of evidence termed by Kuhn (2001 and 2005). Therefore, evidence for the differentiation of higher-order *Argument* skill by the examination in a manner stipulated in the Kuhn's KPI model was elicited, justifying the *Validity Argument (3)*.

Clues for the differentiation of the performance in the formulation of *Arguments* by the examination with reference to the integration of evidence can also be found in the nominal group discussions on the cases in the live script analysis, where examiners' scores differed. A candidate attaining Level 5 was awarded widely different scores among the examiners for Domains 4 and 7 in Paper 2 Question 3 (Table 6.1).

Table 6.1: Marks awarded by the Examiners for Domains 4 and 7

	Domain	
	4 *Synthesis*	7 *Evidence*
E1	2.3	3.6
E2	1	1
E3	1	2.3
E4	3.6	3.6

From Nominal Group Discussion 1, it was found that the discrepancy in marking stemmed from different interpretations of the use of some inappropriate examples to explain whether "soft power is the most effective way for governments to increase their influence in the world". One of the examples the candidate used to illustrate the negative impacts of soft power was a film of the assassination of the North Korean leader, Kim Jong-un. However, s/he did not explain how the film constituted an example of a negative impact of soft power. Both E1 and E2 scored the answer low for Domain 4 *Synthesis* (Table 6.1). E1 deemed it "unreasonable" to argue that the film led to a deterioration of the relationship between North Korea and the US, while the latter considered this example not "significant" or "convincing". Nevertheless, E4 awarded a higher score to it for both Domains 4 and 7 because he "appreciated the use of examples" in the whole answer, despite some of them being inappropriate, in support of the argument for the impact of soft power (Nominal Group Discussion 1). The discussion illustrated that being able to identify and integrate appropriate examples (though the appropriateness may be subject to interpretation) to justify the arguments had been taken into consideration by the examiners in differentiating high-level performance in argument-formulation. Had the candidate been able to integrate appropriate examples as evidence for his/her arguments throughout the answer, the discrepancy in scoring may not have been found.

E1: "…For example, he wrote, 'Third, as there is a movie about the assassination of Kim Jong-un, the relationship between North Korea and other countries is going to be poorer.' I think that is not reasonable. The relationship will not worsen just because of one or some movies. He wrote, 'the impact is far-reaching'. But why is it a negative impact?"

E2: "…. Grids 4 and 5 did not score high because he overgeneralised. The example of Kim Jong-un showed his prior knowledge. It is not irrelevant. But the example was not significant. I awarded B for Grid 1. Paper 2 required prior knowledge, though you may consider the example of assassination of Kim Jong-un extreme. It cannot be applied to all. It is not convincing…."

E4: "He tried to make some arguments. We, adults, may not consider these examples appropriate. But he tried to make use of some examples to formulate arguments and generalise. Though some of the examples were not appropriate, I appreciate the use of examples to formulate arguments."

(Nominal Group Discussion 1)

The evidence for the assessment of the *Integration of Evidence* in *Argument-formulation* in 2015 LS examination can also be complemented by the quantitative analysis in Section 5.1. It was found that the mean score for Domain 7 *Evidence* for Level 5 candidates was 4.257 (Table 5.2), indicating that they were able to "conceptualise evidence" (as described in the Scoring Grid for 5 points of Domain 7. *Evidence*) in formulating arguments. Further evidence for candidates' ability to perform a high level of reasoning as postulated by Kuhn (2001) could also be elicited from the nominal group discussions. The examiners explained that A (5 points) was awarded for Domain 7 in the live script study if the student was able to integrate evidence in the arguments. In the discussion of the scores awarded to some scripts, E1 and E2 mentioned the need to make use of evidence (i.e. the cases given in the sources, examples of impacts, the situation of China) to explain the reasons for undergoing plastic surgery, the positive and negative impacts of a global problem and the enhancement of national power in the examination. E3 also commented that merely quoting some examples, without using examples in formulating argument, cannot fulfil

the requirement of a high-level reasoning.

> *E2: "The previous one just describes the cases in the source. For this one, he tried to make use of the cases in the source to explain (the reasons for undergoing plastic surgery) with concepts, such as self-esteem and the value of appearance."*
>
> *…*
>
> *E1: "… But he failed to explain the impacts. He tried to show that he knows something and tried to use 'evidence'. However, the key word in the question is 'impact'. He needs to explain the impact, no matter whether it is positive or negative. Why is it negative/ positive? This is the assessment of logic. But he was not able to perform it."*
>
> *…*
>
> *E1: "In Part (b) (of Paper 2 Question 1), … He mentioned about the situation in China. But most of it was a factual description. He did not explain the relationship between hard power and the world influence. He had to make use of the examples to explain how the national power will be raised."*
>
> *…*
>
> *F1: "All of you agree that A could not be awarded to Domain 7. Why then?"*
>
> *E3: "The worst example was Singapore. What was the point that he wanted to get to with the use of Singapore as an example?"*
>
> *(Nominal Group Discussion 3)*

It would be more appropriate to employ the wordings of Kuhn (2001) to lay down the criteria for high-level reasoning on the Level Descriptors. Instead of describing a Level 5 candidate as being able to "conceptualise evidence and show respect for evidence", it would be more explicit and concrete to describe the performance as being able to "integrate evidence".

The analysis of *Argument-formulation* with reference to Kuhn (2001, 2005) also helps explain the apparent discrepancy between the Assessment Objectives and the Level Descriptors in terms of "Respect for Evidence" identified in Chapter 4. In view of the need for integrating evidence in demonstrating a high-level of *Argument-formulation*, the demand of the examination might have

been subsumed in Assessment Objective *(h)*, which is about the making of "sound judgements":

> *"(h) to analyse issues (including their moral and social implications), solve problems, make sound judgments and conclusions and provide suggestions, using multiple perspectives, creativity and appropriate thinking skills;"*
> *(HKEAA, 2017)*

A "sound judgement", according to Kuhn (2001, 2005), comprises well-integrated evidential support. Along this line of thinking, there is no discrepancy in the demand of the Assessment Objectives and the Level Descriptors. If the Level Descriptors discern the requirements for high-level performance on judgement-making with regard to the use of evidence, it will be easier for interpretation and aligning more explicitly with the Assessment Objectives.

6.3 Dispositions

The think-aloud protocols (Appendix IV) also illustrated how *Dispositions* (*Knowledge* and *Cultures/Values*) were involved in formulating *Arguments* as postulated by Kuhn (2001, 2005). Examples of the use of higher-levels of *Knowledge*, *Generalisations* and *Principles* (Marzano et al., 2008), in formulating arguments were found in Cases 4 and 8 (Level 5 and 5* respectively). First, Case 4 explained how "cultural conflicts" may arise by a *Generalisation* of Hong Kong people being "discontented with the mainlanders" (Line 53) who peed into a plastic bag (Line 45).

Case 4

44	*some mainland tourists visited the Golden Bauhinia Square*
45	*a mother helped a child to pee into a plastic bag in public places*
46	*maybe this is common on the mainland*
	…
48	*in Hong Kong people may think that this should be done in the toilet*
49	*it is not appropriate to do it in public places*
	…
53	*Hong Kong people will be discontented with the mainlanders*

(Table AIV-4, Appendix IV)

Case 8 (Level 5*) explained the global problems arising from global warming by *Principles* (Lines

67 to 69), which is a higher-level of *Knowledge*. S/he also assessed the scale of the impact of global warming by his/her knowledge of the relationship between carbon dioxide and global problems (Lines 63 to 67).

Case 8

63	*we need to explain why this is global*
64	*carbon dioxide can be a local pollutant*
65	*that is air pollution*
66	*need to talk about global concerns by referring to the effect in the world*
67	*e.g. not only in local*
68	*but in the world, e.g. rise in sea level and so on*
69	*people in the world will be affected*
70	*as so this is a global concern*

(Table AIV-8, Appendix IV)

In the post-interview with Case 6 (Level 5*), some hints on the mental processes involved in *Knowledge Utilisation* can be elicited. The candidate associated the examples of tourists' behaviours with the "travellers of parallel goods" in Hong Kong (Line 57) from his/her own daily experiences and commented on the effect of conflicts on "social harmony" (Line 61), thus justifying it as a global concern.

Case 6

	(How did you go from conflicts to harmony?)
56	*I was thinking about Hong Kong*
57	*in Hong Kong, there are travellers with parallel goods*
58	*it's something similar*
59	*mainland tourists coming to Hong Kong behave in manner different from Hong Kong people*
60	*and then they (Hong Kong people) may have some reactions, and conflicts*
61	*as I have seen, it affects social harmony*

(Table AIV-6, Appendix IV)

In fact, "global warming", "cultural conflicts" and "social harmony" are all in the Curriculum and Assessment Guide (the Curriculum Development Council and HKEAA, 2014). Therefore, candidates might have learnt them in class. When answering this question, they might have associated the data with their experiences and knowledge of concepts from the LS study and then made use of these to formulate arguments. More able candidates made use of higher-level

Knowledge, including *Generalisations* and *Principles*, as categorised by Marzano et al. (2008) in the New Taxonomy.

The use of *Knowledge* and *Cultures/Values* in argument-formulation can also be analysed in terms of the strong and weak methods, which are domain-specific and domain-general respectively, suggested by Klahr and Dunbar (1988) (as cited in Leighton & Gierl, 2011). The use of knowledge of "global warming", "cultural conflicts" and "social harmony", which is part of the curriculum content, involves the domain-specific method. Conversely, drawing on experiences of their own, for instance, the behaviours of mainlanders in Hong Kong in explaining the global concern arising from the increase in international tourism in Cases 4 and 6 (Levels 5 and 5* respectively) is a weak method. They might not have learnt the concerns about international tourism in LS classes. However, in the terms of Posner et al. (1982), candidates "accommodated" their own experiences with the issue and deployed these as examples to a new situation to justify the global concern, showing the mastery of a weak method in the thinking process, in other words, domain-general (not subject specific), according to Klahr and Dunbar (1988).

In contrast with the high-level performance, the protocol of Case 2 (Level 3) did not show the weak method and accommodation. The candidate merely identified the term, "cultural conflict" as the global concern (Line 8). However, s/he did not provide further elaboration on the reason why it deserves global concern. As mentioned in the post-interview, s/he admitted recalling some terms s/he learnt in tutorial classes (Line 19), rather than applying these terms in answering the question. In other words, the think-aloud protocol of Case 2 did not show the command of the strong/ weak method or accommodation of the concepts s/he had learnt. Comparing with Cases 4 and 6 (Levels 5 and 5* respectively), the capability of Case 2 in applying Dispositions, one of the keys to formulating arguments, was weaker. This indicated the LS Examination distinguished

candidates' performance in terms of higher-order thinking skills.

Case 2

6	*need to refer to Source B and Source C*
7	*the most obvious increase is found in transport*
8	*the whole source C is about cultural conflicts*
9	*ie. when asked about global concern, I think these are correct*
10	*how to relate to the tourist arrivals*
11	*because there are more people*
12	*and then…*
13	*resulting in these two global concerns*
	…
	(How did you relate the increase in no. of people and the problem?)
17	*it is found from source B*
18	*When I see CO_2 emission, I feel that it is related to air pollution*
19	*maybe related to tutorial schools*
20	*Source C is about the cultural difference between the mainland and the foreign countries*
21	*I treat it as a global concern*
22	*and then…*
23	*cultural conflicts*

(Table AIV-2, Appendix IV)

6.4 Meta-level Thinking

Besides *Dispositions*, *Meta-level thinking* was also deployed in the thinking processes of argument formulation. Among the 10 cases, only those attaining Level 5 or above (Cases 4, 6, 7 and 8) showed *Meta-level thinking* in the protocols. This is in line with the New Taxonomy (Marzano et al., 2008) which posited that *Meta-level thinking* is a level of cognitive process more complex than *Knowledge Utilisation*. In other words, only more able candidates are expected to master this skill.

Monitoring Accuracy, *Monitoring Clarity* and *Process Monitoring* were the *Meta-level thinking* processes adopted by these four cases. Case 4 (Level 5) checked the accuracy of the analysis of the relationship between the data on tourists and carbon dioxide emissions (Lines 19 to 24). Case 6 (Level 5*) monitored the clarity of the answer by planning for the paragraphing (Lines 44 to

46).

Case 4

19	*the previous point*
20	*besides comparing the two years, the total has increased a lot*
21	*matching what is found in A*
22	*more and more tourists*
23	*the other one should also be discussed from this*
24	*more international tourists*

(Table AIV-4, Appendix IV)

Case 6

44	*and then explain*
45	*after explaining, the paragraph will end and I will open a new paragraph*
46	*the new paragraph is about the social aspect*

(Table AIV-6, Appendix IV)

Process Monitoring by reviewing the fulfilment of the objectives and the logicality of the arguments was evident in the protocols also. All four cases self-evaluated the quality of their answers by reviewing whether the focus was on the "global concern" or not. Case 6 (Level 5*) distinguished the fact of a surge in carbon dioxide emissions from the argument that the problems of global warming leads to a global concern (Lines 41 to 43). S/he then redirected the answer from the description of the carbon dioxide emissions to the concerns over global warming, which was the focus of the question. Cases 4, 7 and 8 (Levels 5 or above) checked the scale of the impacts of "cultural conflicts" (Lines 54 to 57, Case 4), "destruction to tourist spots" (Lines 12 to 14, Case 7) and that brought about by carbon dioxide emissions (Lines 60 to 68, Case 8) to ensure that they were on the right track to explain global concerns.

Case 4

54	*just explained one of the examples*
55	*going back to the question*
56	*it is about global concerns*
57	*one of them, relating to it, is cultural conflicts*

(Table AIV-4, Appendix IV)

Case 6

41	*the increase in CO_2 is not yet relating to the global concern*
42	*more CO_2 is the fact*
43	*the concern maybe about the problems brought about*

(Table AIV-6, Appendix IV)

Case 7

10	*in Source C, there is a destruction of tourist spots*
11	*for example, maybe the natural scenery*
12	*I need to be careful not to take the regional perspective*
13	*adopt a global sense*
14	*the problem happens globally*

(Table AIV-7, Appendix IV)

Case 8

60	*causing the greenhouse effect*
61	*then about the concern over this*
62	*it says two global concerns*
63	*we need to explain why this is global*
64	*carbon dioxide can be a local pollutant*
65	*that is air pollution*
66	*need to talk about global concerns by referring to the effect in the world*
67	*e.g. not only in local*
68	*but in the world, e.g. rise in sea level and so on*

(Table AIV-8, Appendix IV)

To review the logicality, Case 4 assessed the relationships between "the causes and the influences" (Lines 65 to 67 below). Case 8 reviewed the logicality of putting forward "civilisation" as the global concern (Lines 84 to 92) and came up with a more appropriate issue that deserves global concern, "respect for culture".

Case 4

65	*in fact, I think that those parts of my answer in the middle*
66	*the causes and influences are not related together well*
	(What you said just now is an initial line of thinking? If you are really about to write it down, what will you do?)
67	*how to relate them?*

(Table AIV-4, Appendix IV)

Case 8

84	*these are something related to civilization…*
85	*not about civilization…*
86	*but just paying respect to the locals*
87	*as you are visiting that place, you need to do these*
88	*this may reflect…*
89	*but this is not really the concern*
90	*this…*
91	*civilization…*
92	*the concerns should be respect for culture and regions and so on*

(Table AIV-8, Appendix IV)

The process monitoring strategy deployed by Cases 4, 6, 7 and 8 demonstrated the skills in the *Enquiry Phase* of the KPI model also. Even though an entire enquiry process is not required in an examination setting, more able candidates "identified a purpose of the activity" (Kuhn, 2005, p.84), which is the gist of the examination question. All these four candidates scrutinised whether they were able to fulfil the objective of the task by focusing on the explanation of global concerns. They demonstrated the most effective strategy in the *Enquiry Phase* in the KPI model (Kuhn, 2005) by "finding out if X makes a difference in the outcome" (Kuhn, 2005, p.85). X in this task was the assessment of the scale of the impact. Higher-order thinking skills are exhibited in the assessment of the scale of the impact of "global warming" and "respect for culture" and the justification of whether they deserve global concern.

In contrast, Case 2 made no attempt to review his/her answer. S/he simply analysed the data and used some terms, "air pollution" and "cultural difference", to describe the problem shown in the data. In Kuhn's (2005) terms, s/he merely "generates outcomes" (p.140), by composing an answer

with the data, which is the lowest level of *Enquiry*. As s/he did not assess the scale of the impact of the problems s/he suggested, s/he failed to justify how these arouse global concerns, missing the gist of the question.

Case 2

17	*it is found from source B*
18	*When I see CO2 emission, I feel that it is related to air pollution*
19	*maybe related to tutorial schools*
20	*Source C is about the cultural difference between the mainland and the foreign countries*
21	*I treat it as a global concern*
22	*and then…*
23	*cultural conflicts*

(Table AIV-2, Appendix IV)

6.5 The Higher-order Thinking Processes as Illustrated by the KPI Model

The sequence of higher-order cognitive processes involved in answering an LS data-response question in an examination can be summarised based on the KPI model of Kuhn (2001, 2005). Although this model was originally devised for use in an enquiry study of a scientific investigation, it provided a framework for examining the sequential cognitive processes of an issue enquiry in an examination setting. Referring to Dewey's (1937) definition of enquiry, which still stands even after decades, an enquiry is a "transformation of a puzzling indeterminate situation into one that is sufficiently unified to warranted assertion" (Ormerod, 2006, p.900). In an examination setting, the "puzzling indeterminate situation" was structured in the questions, rather than being formulated by a scientific researcher. As the task of a scientific enquiry and an examination question are not exactly equivalent, adaptations of some of the thinking processes in an enquiry are necessary. Owing to the fact that questions were given in the examination, candidates did not formulate enquiry questions. Furthermore, unlike what is suggested in the KPI model, the *Disposition* of candidates did not refer to their knowledge and views on the values of the enquiry.

Instead, they made use of relevant *Knowledge* and *Cultures/Values* in their analyses and arguments.

The sequence of higher-order thinking processes as found from the analysis of the think-aloud protocols can be illustrated by an adaptation of the Kuhn's KPI model (2001 and 2005) in Figure 6.2. Based on the think-aloud protocols, the foremost strategy adopted was an analysis of the data provided in the question. Candidates generalised from the sources the trends of carbon dioxide emissions and the problems stemming from the tourists' behaviours, which were then related to the increase in tourists as shown in the sources. Higher-level performance on *Information-handling* was characterised by a much clearer description of the *Generalisations* by using the data provided, which is in line with the findings in the live script study discussed in Chapter 5.

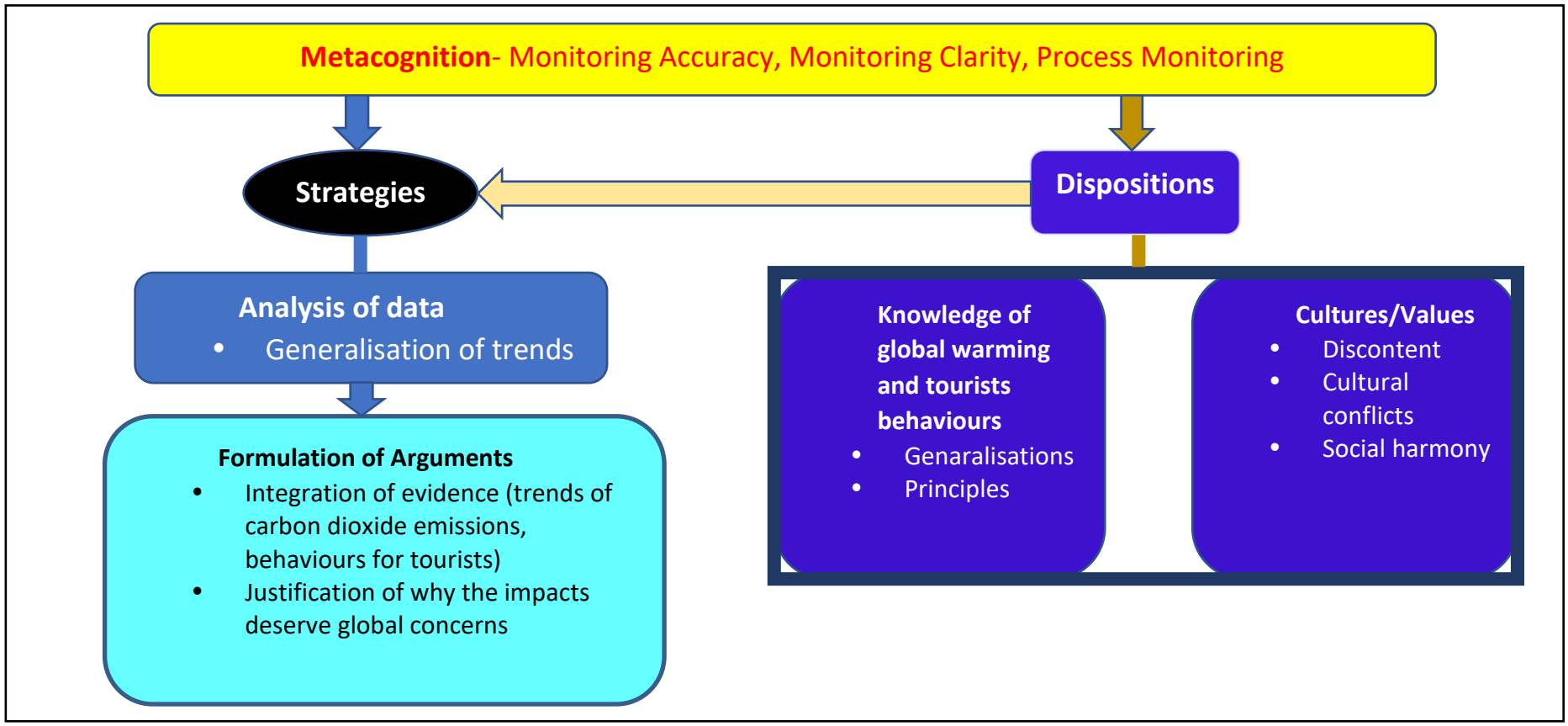
Subsequently, they integrated the trends of carbon dioxide emissions and tourists' behaviours as evidence to support their *Argument* of "global warming" and "cultural conflicts"/ "social harmony" as the global concerns, fulfilling the high-level performance of inferring and reasoning in a manner purported by Kuhn (2001). The think-aloud protocols of Candidates 4, 6, 7 and 8 and Candidate A (Levels 5 or above) in the live script study illustrated the deployment of examples of tourists' behaviours from their own knowledge to explain how the increasing trends in tourists might arouse cultural conflicts. To fully fulfil the requirements of Paper 1 Question 3(b), Candidate A assessed the scale of the impact and explained how the problems deserved global concern.

Values and *Knowledge Dispositions* were *Utilised* (*Knowledge-utilisation* in the terms of Marzano et al. (2008)) in the execution of the strategies of *Information-handling* and *Formulation of Arguments*. Guided by the both the weak and strong methods of thinking as suggested by Klahr and Dunbar (1988) (as cited in Leighton & Gierl, 2011), the candidates accommodated knowledge

they had learnt in the curriculum (domain specific) and from their experiences (domain general). They applied high-level knowledge, including the *Generalisation* of Hong Kong people's view on the behaviours of tourists and the *Principles* of the relationships between carbon dioxide emissions and problems of global warming; as well as that between tourists' behaviours and "social harmony" or "cultural conflicts".

According to Marzano et al. (2008), *Meta-level* skills hold a higher rank than *Knowledge Utilisation* in the New Taxonomy. As an overarching strategy, *Meta-level* skills monitored the performance of this series of thinking processes in conjuring up an answer to the examination question. They reflected on the accuracy, the clarity, the fulfilment of the objectives and the logicality of their answers and subsequently made improvements for their answers.

Figure 6.2 The high-level thinking processes for answering Paper 1 Question 3(b) (Based on the KPI Model (Kuhn, 2001, 2005))



6.6 Chapter Summary

In this chapter, evidence for higher-order thinking processes was gathered from the think-aloud protocols and the thematic analysis of live scripts in support of the Validity Argument *(3) The 2015 LS Examination assesses the higher-order thinking skills of candidates specified in the Level Descriptors.* Candidates attaining Level 5 or above demonstrated more complex *Information-handling* skills: *Generalisation*, as well as higher-level *Dispositions*: *Generalisations* and *Principles*, in accordance with the New Taxonomy (Marzano et al., 2008). The findings from the data analysis and *Dispositions* of related *Knowledge* and *Values* were integrated in the formulation of arguments in response to the questions, fulfilling the criteria for high-level reasoning as put forward by Kuhn (2001, 2005). The highest level of cognitive skill in the New Taxonomy, *Meta-level thinking*, was deployed by these candidates to review the accuracy of the analysis, the clarity, logicality and structure of the arguments.

The findings from my study are in contrast to L. S. Leung's (2017), which was conducted via surveys on teachers, students and policy-makers and classroom observations. L. S. Leung (2017) concluded that LS failed to nurture students with 21st Century skills and the examinations make students "bypass critical thinking", "produce seemingly sensible judgements…without considering fundamental reasoning" and "think restrictively" rather than practising "independent critical thinking" (L.S Leung (2017) Slide 25).

With reference to Bloom (1956), critical thinking involves the intellectual abilities to apply knowledge to analyse new situations and to deal with new problem. Candidates obtaining higher Levels of Performance (Levels 5 or above), Cases 4, 6, 7 and 8 in the think-aloud study and Candidate A in the live script study, were able to demonstrate critical thinking skills as they applied higher-order thinking skills to "new problems" in examination questions. Even though a

generalisation of the performance to all candidates attaining high Levels of Performance is impossible from this study, the live scripts and the think-aloud protocols at least provided evidence for higher-order thinking skills, which is in contrast with the findings of L. S. Leung (2017). Accommodation of prior knowledge was demonstrated in the reasoning and justification of arguments in response to "new problems" in the examinations. Even though these candidates adopted similar thinking processes and were "restricted" by the predetermined "enquiry questions" in the examination, the *Dispositions*, the ways of analysis and the arguments formulated were different, indicating an independent thinking process. "Formulating enquiry questions" is just one of the higher-order thinking skills according to Corliss & Linn (2011) and the KPI Model (Kuhn, 2001, 2005). The elimination of the assessment of this skill in the written examination did not undermine the performance of other higher-order thinking skills. Furthermore, candidates who merely associated the questions with some terms in their memory might not be able to formulate relevant and coherent arguments, like Case 2 (Level 3) in the think-aloud study.

As discussed in Chapter 5, being able to differentiate candidates' performance in accordance with the level of complexity as stipulated by cognitive models, as well as assessing higher-order thinking skills, are pieces of evidence for the appropriateness of the examination. Although some teachers and students might not perceive that LS nurtures critical thinking as shown by L. S. Leung (2017), critical thinking was evident in the authentic performance of candidates in the examination in my study. This may suggest a discrepancy between the understanding of teachers and students on the learning of critical thinking and the performance requirements in LS, which may be worth further research.

The applicability of the process of validity evaluation on a large-scale examination illustrated in Chapters 4, 5 and 6 will be discussed in the following chapter.

CHAPTER 7 APPLICABILITY OF THE VALIDITY EVALUATION PROCESS

In Chapters 4, 5 and 6, the content and substantive validity of the 2015 LS examination was evaluated by using the Argument-based Approach postulated by Kane (2013, 2015). The validation process adopted was based on both quantitative and qualitative evidence drawn from multiple sources: a content analysis, a live script study, nominal group discussions and a think-aloud study. Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the New Taxonomy (Marzano et al., 2008) and Kuhn's (2001, 2005) KPI model were deployed as the analytical framework for the substantive validity evaluation. In this chapter, the implications, limitations and factors contributing to the applicability of the validity evaluation process adopted in this study will be discussed. Even though part of the live script study, the nominal group discussions and the think-aloud study were secondary data, as this study focuses on the methodological aspects of the assessment validation process, the methodology for gathering the secondary data set will also be evaluated in this chapter.

7.1 Implications of the Validation Process

The validation process adopted by this study has shown how evidence can be gathered from multiple sources and dual perspectives (for evaluating content and substantive validity). In this study, the quantitative evidence that most validation studies rely on was also complemented by qualitative data, which may help to enhance assessment development. These implications will now be discussed in more detail.

7.1.1 Gathering Evidence from Multiple Sources

Firstly, this study illustrated how evidence could be gathered from multiple sources, including the authentic performance of candidates, views from examiners and think-aloud protocols, for the evaluation of the validity of the LS examination. In order to uphold the accountability of this high-stakes public examination for all Secondary 6 students in Hong Kong, an evidence-based validity evaluation is of paramount importance. In this regard, following Kane's (2013, 2015) Argument-based Approach, quantitative and qualitative evidence was drawn from a content analysis of assessment documents, a live script study, nominal group discussions and a think-aloud study to investigate the Validity Arguments of the appropriateness of the Level Descriptors in the differentiation of candidates' performance and the assessment of higher-order thinking skills.

The collection of empirical evidence from multi-faceted sources was advocated by a number of assessment validity theorists, including Goldstein (2015) and Messick (1989), as a means to enhancing the credibility of the assessment validation. In this study, a triangulation of evidence from multitudinous sources was carried out for validating the appropriateness of the interpretation and use of the examination results. This empirical study filled the gap in the assessment validation literature, operationalising a validation process on a large-scale examination, drawing on multiple sources of evidence, which was suggested to be a direction of future research by Shaw et al. (2012).

In this study, both quantitative and qualitative evidence for validity was gathered from three different sources, each complementing one another. While the ANOVA analysis of the live scripts in Sections 5.1 and 5.2 provided evidence for the differentiation of the performance on skill domains of candidates attaining various Levels of Performance, in fact the scores awarded by the examiners according to the Scoring Grid derived from the Level Descriptors cannot provide details of the cognitive skills of candidates in the evaluation of an alignment with cognitive models

(as discussed in Section 5.3). Whether or not the examination differentiated the complexity of skills application by candidates as stipulated in cognitive models could not be shown by the scores awarded. The descriptions on the Scoring Grid merely constitute a guideline for scoring the skill performance, rather than reflecting directly the performance details. Therefore, the qualitative analysis of the live scripts and think-aloud protocols supplemented for this inadequacy of a purely quantitative study. By the interpretive enquiry based on qualitative evidence in this study, other aspects could be examined. These include the complexity of the knowledge and concepts, as well as the analytical skills applied, the variation in the coherence of arguments, the evaluation with reference to assessment criteria and the use of evidence in support of candidates' arguments. As such, an in-depth analysis of the alignment with cognitive models as well as the sequential cognitive processes were made possible by the thematic analysis of the qualitative data from nominal group discussions, the live scripts and the think-aloud protocols.

7.1.2 Gathering Evidence for Content and Substantive Validity

This study gathered evidence for both content and substantive validity, evaluating the validity of the examination from multiple aspects. Fundamentally, for an examination to be valid, the assessment requirements have to fulfil the Assessment Objectives. In this regard, the content analysis of the question papers of the 2015 HKDSE LS Examination, the Level Descriptors and the Assessment Objectives provided evidence for the content validity. However, this kind of content analysis can only be deployed to evaluate whether the examination is designed as expected. To evaluate the substantive validity, in this study, on the one hand, evidence on the differentiation of the actual performance of candidates in accordance with several cognitive models was elicited from the live script study and nominal group discussions of examiners. In addition, the think-aloud study provided further evidence for the command of higher-order thinking skills as

stipulated in the Assessment Objectives and Level Descriptors.

Furthermore, gathering evidence for content and substantive validity can help respond to concerns on the consequential aspect of validity, which is termed an "external" dimension by Shaw et al. (2012, p.171). According to Messick (1995), different types of validity are interconnected and can be unified under construct validity. As such, one type of validity may lead to another. The justification of the "internal" content and substantive validity can enhance the "external" social acceptance of the examination. In this study, although direct evidence on the consequential aspect of the validity of the examination was not available, evidence on the differentiation power and the assessment of higher-order thinking skills (in Chapters 4, 5 and 6) in support of the content and substantive validity, is able to address the social concerns over this new core subject in the HKDSE. In a society with a "culture of testing" (Manns et al., 2018, p.13), stakeholders attach paramount importance to public examinations. Therefore, an evidence-based validation process for addressing the queries or concerns of stakeholders in society is necessary.

Empirical studies of content and substantive validity may provide grounds for decision-making in the review of the curriculum and assessment of a subject in response to the social concerns. Findings from this study on the differentiation of candidates' performance by five Levels of Performance can provide evidence for opposing a change to a pass/fail grading system proposed by a non-binding motion in the Education Panel of the Legislative Council of Hong Kong (Legislative Council Secretariat, 2017). Since this study has shown that the 2015 LS Examination was able to differentiate candidates into five Levels of Performance, the existing grading system is worth maintaining, rather than being simplified to the distinction of "pass/fail". If a "pass/fail" grading system is adopted, candidates' performance cannot be so widely distinguished and high ability candidates cannot be recognised. In Section 5.1, it was found that the 2015 LS Examination

differentiated candidates' overall performance as well as performance in the majority of skill domains at all levels (with the minor exceptions of the differentiation of Domain 5 *Evaluation* and Domain 6 *Cultures/Values* between Levels 1 and 2),. In addition, in Section 5.3.2, the qualitative analysis of live scripts provided evidence for the performance of *Evaluation* by candidates attaining Level 5 only and *Synthesis* by Candidates at Levels 4 and above. If there are only "pass/fail" grades in the examinations, candidates who are capable of the higher-level thinking skills, *Synthesis* and *Evaluation*, cannot be distinguished and will only be awarded the same "pass" as others capable of performing lower order skills as stipulated by Bloom's Taxonomy (1956), such as *Analysis*.

7.1.3 Evidence of Sequential Cognitive Processes from Think-aloud Study

In Chapter 6, the think-aloud study provided data on the sequential cognitive processes and the higher-order metacognitive skills of candidates, including the monitoring of accuracy, clarity and logicality, which cannot be found by an analysis of the live scripts alone. Think-aloud studies have been put forward by assessment scholars, including Pellegrino et al. (2016) and Shaw et al. (2012), as a source of evidence for assessment validity. Ericsson and Simon (1980) and Olson et al. (1984) also considered think-aloud methods as a means to probe into the verbal form of working memory (as cited in Charters, 2003). Without the think-aloud study, the sequence of various cognitive processes cannot be investigated. Nevertheless, validity evaluation based on think-aloud studies is scarce in the literature.

In addition, evidence of the highest-level cognitive process in the New Taxonomy, *Meta-level thinking* (Marzano et al., 2008), was elicited by the think-aloud study in this dissertation. Evidence of reviewing the answers for improving the accuracy, clarity and logicality was identified from

the think-aloud protocols (Section 6.4). While the quantitative and qualitative analyses of the authentic scripts shed light on the final product of the thinking process in answering examination questions, they did not provide hints on the overarching complex cognitive processes behind the scene. The process of reviewing the accuracy, clarity and logicality of the answers before composing them in writing could only be illuminated in the think-aloud study (Section 6.4).

7.1.4 Informing Assessment Development

The proposed validation process may inform test developers of the directions for improvements in test design and the grading process. The use of validity evidence to enhance test development has been advocated by validity theorists. Messick (1989) suggested that assessment validation helps to improve the assessment itself (as cited in Moss et al., 2006). As Moss et al. (2006) put it, "validity is as much an aspect of test development as it is of test evaluation" (p.116). Moss (1996) postulated that assessment validations may adopt a dialectic approach with "analytic reflexivity" (as cited in DeLuca, 2011, p.311). In other words, test developers can reflect upon the findings from the validation so as to bring about assessment enhancement. As illustrated in this study, the validity evaluation lent support for the need to enhance the Level Descriptors of the LS Examination in relation to the application of cultural knowledge and integration of evidence, as well as the performance on *Evaluation*. The thinking process of candidates when answering an examination as found from the think-aloud study provided clues for test developers to assess the appropriateness of the demand of the question in the question setting process.

Areas for further enhancement in the Level Descriptors and candidates' mastery of thinking skills were identified from the validation process in my study. In particular, the inadequacy in the description of the use of *Knowledge* on culture and values on the Level Descriptors (Section

5.3.1.2); the wording for the use of evidence at Level 5 on the Level Descriptors ("conceptualising evidence" instead of "integrating evidence", which was the term used by Kuhn (2001, 2005)) (Section 6.3); and below expectation scores for the *Evaluation* skill, even for candidates attaining the top Level of Performance (Level 5) (Section 5.3.2.1) were identified. In this regard, this multi-sourced validity evaluation process can provide feedback for test developers and teachers to enhance the Level Descriptors as well as the learning of thinking skills.

Besides reflecting on the Level Descriptors, the validation process adopted in this study can also shed light on the consensus-making process in marking and grading. The nominal group discussions and the re-scoring of live scripts (Chapter 5) illuminated how examiners made judgement on the skill performance of candidates. This study showed a consensus among the examiners on the need for coherence in arguments and evaluation with reference to clearly discerned criteria for attaining the top Level of Performance. Though the consensus-making process was not the focus of the study, the nominal group discussions could be further analysed to investigate how to reach consensus more effectively on the criteria for marking and grading, thus enhancing the assessment in these aspects.

Furthermore, the think-aloud study (Chapter 6) allowed test developers to probe into the sequence of cognitive processes and examine whether adequate data were provided in the data-response questions and whether the *Dispositions* required were within the realm of the curriculum. The think-aloud protocols provided clues for the use of the data on international tourists and carbon dioxide emissions by candidates in the specific questions, as well as how the candidates might have associated the data with concepts and knowledge in the curriculum. As illustrated in Section 6.3, candidates adopted *Disposition*s, including "global warming", "cultural conflicts" and "social harmony", which are all in the *Curriculum and Assessment Guide* (the Curriculum Development

Council and HKEAA, 2014). Based on the evidence from the protocols and the post-interviews of this think-aloud study, test developers may evaluate the linkage between the assessment and the curriculum, as well as the appropriateness of the data provided and the *Disposition* requirements of the examination questions. Besides adopting the think-aloud study as a post-mortem validation process after the examination, in the future examination paper development stage, test developers may go through the sequential thinking process shown in the think-aloud protocols (Figure 6.2) in order to scrutinise the usefulness of the data provided and the appropriateness of the *Disposition* requirements.

In fact, "analysis of errors" as advocated by Pellegrino et al. (2001, p.207) could have been conducted on the think-aloud protocols. Based on this sort of analysis, test developers may identify the type of data or the question-wording that candidates have difficulties in interpreting or deploying and pay attention to these in question-setting in the future. However, the present study aimed at evaluating the assessment of higher-order thinking processes based on the think-aloud protocols. As such, common errors of candidates in data analysis were out of the scope.

In this dissertation, the scoring of live scripts also allowed a reflection on the appropriateness of the Level Descriptors in the grading process of the examination. Since the scripts were marked according to the question-specific marking guidelines, whether the Level Descriptors appropriately reflect and differentiate the mastery of skills of candidates cannot be evaluated merely by the marks awarded. In a bid to justify the substantive validity, the appropriateness of the Level Descriptors in the determination of the cut points/scores for different Levels of Performance in the grading process should be evaluated. Therefore, re-scoring and qualitatively analysing the scripts at various Levels of Performance with reference to the Level Descriptors directly provided evidence for the differentiation of cognitive skills by the examination.

7.2 Limitations of the Validation Process

In this section, the limitations of the validation process adopted by this study will be examined with reference to the comprehensiveness of the empirical data, the constraints in re-scoring the live scripts and the think-aloud study.

7.2.1 A Lack of Comprehensiveness

First of all, this study did not aim at a comprehensive evaluation of all the assessment tasks of the LS assessment. Only the validity of the written examination was evaluated. However, according to Shaw et al. (2012), being able to "encompass the whole assessment process" (p.161), from the administration to the impacts of the assessment is important in the validation process. In this study, even though the written examination covered a high proportion (80%) of the public assessment as a whole, due to the omission of the School-based Assessment, the IES (Independent Enquiry Study) (which accounts for the remaining 20%), a unique aspect of the Level Descriptors of the subject on the enquiry skills (as shown in the following) was not evaluated.

> *"• show initiative and self-management skills and reflect comprehensively and systematically throughout the enquiry learning process"*
>
> *(HKEAA, 2014)*

In fact, this evaluation process may also be adapted for investigating holistically the validity of both the written examination and the IES component. For evaluating the differentiation of the Levels of Performance by the IES, the live script study adopted in this research could be modified to study authentic IES reports quantitatively and qualitatively. In this study which focused on the written examination, the whole enquiry process as described by Kuhn's (2001, 2005) KPI model was not investigated. In an evaluation of the validity of the IES, the enquiry skills, not only those skills assessed in the written examinations (*Information-handling, Application of Dispositions* and *Arguments Formulation* in the KPI model), but also the formulation of enquiry questions and the

design of enquiry methodology and enquiry plans, may constitute the scoring grid and may be analysed thematically. As such, the KPI model (Kuhn, 2001, 2005) adopted in this study would also be applicable. According to the KPI model, being able to "find out if X makes a difference in outcome" is the highest-level of *Enquiry*. Therefore, to evaluate the mastery of higher-order *Enquiry* skills by candidates, a thematic analysis of authentic IES reports could be conducted to examine whether the highest-level of *Enquiry* can be performed by candidates.

Nevertheless, not all of the validation process proposed in this research can be conducted on the component of School-based Assessment. As the IES is an extended school-based assessment task, a think-aloud study of the thinking process involved in the whole task, which might have been carried out for more than a year, is not possible. Instead, in-depth interviews or focus groups are more appropriate tools for collecting data from the candidates on the processes and difficulties in formulating enquiry questions and designing enquiry plans.

7.2.2 Difficulties in Re-scoring the Live Scripts

In the nominal group discussions, examiners expressed difficulties in scoring the live scripts by skill domain. In fact, in the actual marking and grading process of the HKDSE LS, examiners have been adopting a holistic, rather than an analytic approach. This may explain why they found it hard to atomise and score the performance for each of the skills. For example, in Nominal Group Discussion 1, examiner E2 commented on the overlapping of the criteria for scoring Domain 4 *Synthesis*, Domain 5 *Evaluation* and Domain 6 *Cultures/Values*.

> *E2: "…There may be some overlapping between Grids 4 and 6. Grid 4 is focused on the logics of the arguments. Should the consideration of counter-arguments be awarded in Grid 6?"*
>
> *…*
>
> *E2: "But (the discussion of) 'counter-arguments' (in Domain 6) might overlap with Domain 5 a bit. In Domain 5, 'one-sided arguments' is also described. Therefore, for one-sided arguments, i.e. without considering counter-arguments, scores will be awarded to Domain 5."*
>
> *…*
>
> *(Nominal Group Discussion 1)*

By nature, *Synthesis*, *Evaluation* and the use of knowledge or concepts of *Cultures/Values* (Domains 4, 5 and 6 on the Scoring Grid) are the component skills of *Argument-formulation*. As discussed in Section 3.4.2, the relationship among these three skills was reflected in the design of the Scoring Grid (Appendix I). They are all grouped under "*Formulation of viewpoints, opinions and suggestions*" on the Scoring Grid. The consideration of counter-arguments, various impacts/viewpoints and cultures/values other than the candidate's own viewpoints and cultures/values is a criterion for formulating logical arguments. As a matter of fact, according to the Scoring Grid, if a candidate merely "gives irrelevant opinions, suggestions and ungrounded arguments", "provides one-sided arguments" or "elaborates on their own views based on their own values/cultures", Ds have been awarded for all the three domains (Domains 4 *Synthesis*, 5 *Evaluation* and 6 *Cultures/Values*). In this regard, these domains are interrelated, though referring to different aspects of an argument.

To enhance the examiners' understanding of the Scoring Grid and the scoring criteria for each of the skill domains, sample scripts attaining Level 5 were deployed in the pre-meeting of the joint study for clarification. In the subsequent nominal group discussions, the examiners, who were familiar with consensus-marking, made adjustments to their marking standards. As a result, an

alignment in the marking standard can then be observed, especially for candidates attaining Levels 5 and above. As shown in Table 7.1, even though there was a variation in the percentages (ranging from 30.78% to 71.96% awarded 5 points), all examiners (except E4, who awarded 3.6 to a slightly higher percentage of scripts) awarded the highest scores on the scale (5 points) to the largest proportion of candidates. Besides the ceiling effect of the 5-point scale, the discussion with sample scripts of Level 5 in the pre-meeting might have played a role in aligning the marking standards already, thus ensuring the validity of the scores as the secondary data sources for my study.

Table 7.1: The percentages of different scores awarded by Examiners to candidates attaining Level 5, 5* and 5**

Examiner	% awarded 5	% awarded 4	% awarded 3.6	% awarded 3	% awarded 2.3	% awarded 2	% awarded 1
E1	**35.29**	18.43	23.53	9.41	10.39	1.18	1.77
E2	**71.96**	5.69	20.20	0.39	1.77	0	0
E3	**32.94**	20.51	23.27	9.66	8.68	1.97	2.96
E4	**30.78**	27.26	**31.37**	4.31	5.49	0.59	0.20

In contrast, the scores for candidates attaining Level 3 varied most among the examiners (Table 7.3). Examiners E3 and E4 were more lenient with candidates attaining Levels 3 and 4, whereas E1 was stricter (Tables 7.2 and 7.3). Both in fact awarded 5 points to a higher percentage of scripts attaining Level 4 than Level 5 scripts. Nevertheless, in comparison with Level 4 scripts, these two examiners awarded 5 points to a much lower percentage of scripts and lower points to a higher percentage of scripts attaining Level 3 (more 2.3 points by E3 and more 3.6 points by E4) (Table 7.3). They concurred with Examiners E1 and E2 on the scoring standard that the Level 3 scripts showed a weaker performance.

Table 7.2: The percentages of different scores awarded by Examiners to candidates attaining Level 4

Examiner	% awarded 5	% awarded 4	% awarded 3.6	% awarded 3	% awarded 2.3	% awarded 2	% awarded 1
E1	5.01	**18.85**	**21.00**	**18.14**	16.23	12.17	8.59
E2	0.48	**44.52**	29.05	1.67	24.29	0	0
E3	**43.33**	19.52	23.33	5.24	8.10	0.48	0
E4	**34.52**	**30**	**32.86**	2.14	0.48	0	0

Table 7.3: The percentages of different scores awarded by Examiners to candidates attaining Level 3

Examiner	% awarded 5	% awarded 4	% awarded 3.6	% awarded 3	% awarded 2.3	% awarded 2	% awarded 1
E1	0	2.67	0	**34.00**	9.33	**36.00**	18.00
E2	19.33	**27.33**	21.33	8.67	23.33	0	0
E3	**28.00**	20.67	15.33	8.67	17.33	4.00	6.00
E4	27.33	26.00	**35.33**	4.67	6.67	0	0

Furthermore, as the scales of scores are not uniform on the Scoring Grid across the domains, varying from four- to five-point scales (though both scales took 1 as the minimum and 5 the maximum), it might be difficult for the examiners to apply the Grid on the performance in the middle of the scale, resulting in a greater variation in scoring standards observed among the examiners for scripts attaining Levels 4 and 3. Examiner E1 was the strictest, scoring lower points (3 and 2 points) for a much higher percentage of scripts attaining Levels 4 and 3. Examiner E3 was on the other end of the scale, being more lenient, awarding 5 points for a higher percentage of scripts at Levels 4 and 3. Nevertheless, the discrepancy in scales on the Scoring Grid was inevitable in this study as the scoring criteria have to be derived from the Level Descriptors for evaluating the appropriateness in differentiating the performance of candidates. Due to the fact that the differentiation between Levels 1 and 2 in terms of the *Formulation of Viewpoints, Opinions and Suggestions*, as well as *Respect for Evidence* was not explicit on the Level Descriptors, a four-point scale was adopted for these skill domains on the Scoring Grid. Though this might have made it difficult for the examiners to apply the same standard in scoring all the

eight domains, especially for the performance in the middle of the scale, 4-point scales have to be adopted according to the Level Descriptors.

To align the marking standard among the examiners, the cases with large discrepancies were discussed and examiners came to a consensus with mark adjustments. In order to improve the re-scoring process, more scripts should have been discussed among examiners in the nominal group meetings.

For the quantitative analysis in Chapter 5, the averages of scores among the four examiners were used. Although statistically significant decreases in the average of the examiners' scores from Level 5 to Level 1 were found for most of the skill domains in this study, to enhance the consensus in the scoring standard among examiners in the validation process, samples attaining various Levels of Performance should be discussed in the pre-meeting.

7.2.3 Variation in the Nature of the Samples of Different Levels of Performance

As the majority of live scripts were deployed from the joint study with the HKAGE, the distribution of scripts attaining different Levels of Performance in this study reflected the performance of the members of the organisation, which concentrated on Level 4. To investigate the differentiation of candidates' performance across the whole spectrum of Levels of Performance, live scripts on the HKEAA website were deployed. In contrast to the live scripts from the joint study, the scripts from the HKEAA website did not provide data of the performance of an individual candidate on the whole written examination. Instead, this was a set of selected typical performances for each question in the examination, with less variation in performance across questions. To facilitate a comparison of the findings and to preserve authenticity, it would

have been more desirable for the live scripts to be of the same nature, for instance, being the whole scripts of individual candidates attaining various Levels of Performance.

7.2.4 Design of the Think-aloud Study

Constraints of this validation process may have stemmed from the design of the think-aloud study also. The retrospective nature of the think-aloud study, which served as the secondary source of data in the validation process discussed in Chapter 6, might have led to a loss in traces of the actual thinking process involved in the examination setting. The thinking process deployed in formulating a thorough argument as well as putting it in writing may not have been examined. For instance, even though Candidate 8 was awarded Level 5*, his/her protocol was fragmented and incoherent. Only one example, "rise in sea level" (Line 68) was cited to illustrate the global impact of global warming.

Case 8

63	*we need to explain why this is global*
64	*carbon dioxide can be a local pollutant*
65	*that is air pollution*
66	*need to talk about global concerns by referring to the effect in the world*
67	*e.g. not only in local*
68	*but in the world, e.g. rise in sea level and so on*
69	*people in the world will be affected*
70	*as so this is a global concern*

(Table AIV-8, Appendix IV)

While conducting a think-aloud study in a live examination is impossible, to illustrate what participants need to do in the retrospective think-aloud study, the researcher in the joint study demonstrated how to think aloud by working on Part (a) of the same question before the recording of the think-aloud protocols by the participants. To further enhance the quality of the think-aloud protocols in future studies, participants may need to try out the process with another question.

Feedback from the researchers may help them understand the task better and then think aloud for the whole process of answering a question in an examination.

In spite of the demonstrations and trials prior to a retrospective think-aloud task, alterations to the thinking process in a second attempt of the examination question are inevitable. Firstly, they might merely retrieve their answers rather than thinking over the questions as in an actual examination. Some participants might have got some answering points for the question from their teachers or peers after the examination. As such, the investigation of the use of *Dispositions* in *Arguments* might not reflect what they actually did in the examination.

In addition, some of the *Dispositions* that can be used in answering the question in the think-aloud study might be out of the participants' memories after moving onto another field of study half a year from the examination. *Dispositions* such as the global problems related to global warming might be lost. Therefore, they might not be able to cite any examples of problems of global warming in the justification of a global concern and thus fail to achieve the highest level of reasoning by integrating evidence to formulate *Arguments* as suggested by Kuhn (2001, 2005). To better reflect the thinking process adopted in the examination, a think-aloud study shortly after the examination may be more appropriate.

Even if the sequential thinking processes adopted in the examination was revealed in the retrospective think-aloud protocols in my study, the processes may have varied among candidates (Someren et al., 1994). Although all candidates of Levels 5 or above deployed *Meta-cognitive skills* as found in this study, there is no evidence that candidates who did not deploy these skills are of lower ability. Instead of reviewing the logicality or clarity of the answers, it is worth investigating whether some candidates make the answers logical and clear in the first attempt in

answering the question, without the need to look back. Furthermore, there may be variations in the mental processes for triggering and incorporating *Dispositions* in *Arguments*. As shown in the post-interview with Case 7, a Level 5** candidate, s/he did not have any intention to apply "concepts" in answering the question (Line 113) even though s/he had formulated arguments with "global warming". S/he did not consider "global warming" a "concept". More empirical data are necessary to further examine these variations in the mental processes of *Dispositions and Arguments*.

Case 7

	(Did you answer this question with anything you have learnt in class, for example, concepts?)
113	*I don't see the need in using concepts in this question*
114	*for example, some questions may be about quality of life*
115	*then we need to talk about concepts*
116	*but there is no need for this one*
	…
	(Did you think of concepts that are related to the question?)
125	*Seldom*

(Table AIV-7, Appendix IV)

7.3 Factors Affecting the Applicability of the Validation Process

In light of the findings from the validity evaluation on the 2015 HKDSE LS Examination, the implications and limitations of the validation process as discussed in the previous two sections, the factors determining the transferability of the methodology of the validation process deployed in this study will be analysed in this section. Here how the nature of the assessment and grading mechanism, as well as the availability of experienced examiners determine the applicability of the validation process adopting the Argument-based Approach of Kane (2013, 2015), which comprises a mixed quantitative and qualitative live script study and a think-aloud study, will be discussed.

7.3.1 Standards-referenced Grading Mechanism

As discussed in Section 7.1.1, the validation process is characterised by sourcing evidence from both qualitative and quantitative data, including the scores, live scripts, nominal group discussions and think-aloud studies. The process for gathering these data may vary according to the grading mechanism.

The grading mechanism is a factor determining whether this validation process is transferable or not. In the HKDSE LS Examination, the Levels of Performance of candidates are graded by standards-referencing. After the marking of scripts according to question-specific marking guidelines, examiners determine the threshold scores for various Levels of Performance with reference to the authentic performance of the candidates and the Level Descriptors. As illustrated in Chapter 5, to evaluate the appropriateness of the Level Descriptors, live scripts were re-scored with respect to a Scoring Grid derived from the Level Descriptors.

However, if the marking guidelines of the examination questions are standards-referenced by which the marks awarded in the marking stage are converted to the grades of the subject directly, or if the grading is based on norm-referencing, re-scoring the live scripts merely yields evidence for the marking process, rather than the grading process. Alternatively, for these cases, the marks of the examination could be analysed quantitatively with an external criterion measure for evaluating the criterion validity, as exemplified by the correlation analysis of Pellegrino (2016). To gather evidence from multiple sources for triangulation in the substantive validity evaluation, a qualitative analysis of authentic scripts and think-aloud protocols as shown in Chapters 5 and 6 can also be viable for investigating the alignment of the assessment with cognitive models.

7.3.2 Nature of the Assessment Domains

As discussed in Section 7.1.2, the validation process in this study illustrated an evaluation of both the content and substantive validity on the LS Examination from multiple sources of evidence (in Chapters 4, 5 and 6). To justify the substantive validity, the alignment between the differentiation of performance by the assessment and that stipulated by cognitive models was examined. In this regard, the cognitive model(s) to be deployed hinge(s) upon the nature of the assessment domains: whether it is skill- or content-based.

Chapter 5 showed that for evaluating the LS Examination, a skill-based examination, cognitive models mapping out the relative cognitive demands of thinking skills, Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson and Krathwohl, 2001) and the New Taxonomy (Marzano et al., 2008) are appropriate. As specified in the *Curriculum and Assessment Guide of LS* (HKEAA, 2014), candidates are assessed on their application of thinking skills for analysis and discussion, as well as their judgement-making in relation to contemporary issues. The examination was found to differentiate candidates' performance in alignment with the taxonomy of complexity of cognitive skills as stipulated in the cognitive models. Also, as judgment-making is one of the key assessment objectives in LS, Kuhn's (2001, 2005) KPI model was deployed in Chapter 6 and it was found that the higher-order thinking process, including argument formulation, was assessed.

Nevertheless, for assessments which are content-based, models placing more emphasis on the understanding and application of the subject-specific knowledge and concepts allow the checking of the appropriateness of the differentiation in the performance. For these cases, the alignment with the Knowledge Domains of Bloom's Taxonomy (1956) and the New Taxonomy (Marzano et al., 2008) as shown in Section 5.3.1 could take up a more significant role in the evaluation of substantive validity.

7.3.3 Trained Examiners

As shown in Section 7.2.2, the examiners found it difficult to score the scripts according to the Scoring Grid derived from the Level Descriptors and to align the scoring standard especially for Levels 3 and 4. Since the proposed validation process in this study comprises a re-scoring process of the live scripts with reference to a Scoring Grid derived from the Level Descriptors, the availability of examiners who are familiar with the Level Descriptors is a pre-requisite. Shaw & Imam (2013) and Cook et al. (2016) agree on the significance of training examiners for assessment validation processes. The examiners must have a clear understanding of the requirements stipulated by the Level Descriptors of each Level of Performance. This minimises the possibility of a poor differentiation of the thinking skills due to the inability of examiners to score the scripts according to the Level Descriptors.

In the joint study, which provided a set of secondary data for my study, examiners who had experience in the grading process were selected to facilitate the consensus-making process in the re-scoring process. A pre-meeting to familiarise the examiners with the use of the Scoring Grid and to align the marking standard was conducted with an aim of building up a "training group effect" as discussed by Baird et al. (2017). The pre-meeting could have cultivated a "group culture" (Baird et al., 2017) by the discussions of the application of the Scoring Grid on some sample scripts before the re-scoring process, facilitating the achievement of a consensus in the marking standard, in particular for the Level 5 performance as shown in Section 7.2.2. Referring to the extract of Nominal Group Discussion 1 below, examiners E1 and E2 had consensus in awarding As (5 points) to Domain 3 for scripts which considered multiple perspectives. This might have been the result of the "group culture" of a consensus on the scoring criteria established in the pre-meeting, providing further support for having trained examiners to enhance the quality of the data for the validation process

E1: "... he was able to point out the possibility for the government to limit the diversity of views on some issues in the media. Then he continued with the explanation of 'the determination of the citizens', which was about the support from the citizens on maintaining press freedom in Hong Kong. It was then followed by 'law' and 'pressure from the business sector' as factors. In Part (s), he explained the factors influencing press freedom from various perspectives. ..."

E2: "He considered many different perspectives."

(Nominal Group Discussion 1)

7.4 Chapter Summary

In a nutshell, the validation processes adopted in the evaluation of the 2015 HKDSE LS Examination, which made use of a live script study, nominal group discussions and a think-aloud study, provided multiple-sourced evidence for content and substantive validity. From the validity evaluation process, the "appropriateness for the interpretation and the use" of the examination results could be enhanced. As illustrated in Section 5.3.1.2, to reflect more appropriately the actual skill performance of candidates and the requirements of the examination as stipulated in the Assessment Objectives, considerations of values and cultures should be added to the Level Descriptors. Furthermore, as discussed in Section 6.2, to depict the performance of the Level 5 candidates, "*integration of evidence",* which is the highest-level of reasoning as defined by Kuhn (2001), should be spelled out in the Level Descriptors.

However, to apply this validation process, the assessment should possess several characteristics. Firstly, the assessment should be skill-based for deploying the cognitive domains of Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001) and the New Taxonomy (Marzano et al., 2008) as the analytical framework. As for content-based assessments, cognitive models placing emphasis on the differentiation of content-related *Dispositions* will be

more appropriate as the analytical framework. Secondly, the validation process is meaningful for standards-referenced, rather than norm-referenced assessments. In addition, experienced examiners are indispensable for the live script study in this validation process.

To further enhance this validation process, a comprehensive evaluation should be conducted on all the components of an examination, which means including the IES[54] in the study of LS Examinations. The quality of the data can also be improved by building up consensus among examiners in the scoring process of the live scripts and deploying live scripts of similar nature across the Levels of Performance. Nevertheless, the possibility of alterations to the thinking process or loss of information due to the retrospective nature of think-aloud studies and the individual variations in cognitive processes remain issues of the proposed validation processes for further exploration.

[54] *IES (Independent Enquiry Study) is the School-based assessment task in the HKDSE LS Assessment.*

CHAPTER 8 CONCLUSION

In view of the inclination of literature on assessment validity towards the theoretical aspects, this study operationalised a process for evaluating the validity of large-scale high-stakes assessments, as exemplified by the validation of the 2015 HKDSE LS Examination.

Based on the definition of assessment validity by Messick (1995) and the argument-based Approach of Kane (2013, 2015), a multi-faceted evaluation of the content and substantive validity of the 2015 HKDSE LS Examination was conducted for justifying the following Validity Arguments:

> *(1) The Assessment Objectives and the assessment criteria of the 2015 HKDSE LS Examination align with the Level Descriptors;*
>
> *(2) The Level Descriptors appropriately differentiate the performance of candidates;*
>
> *(3) The 2015 LS Examination assesses the higher-order thinking skills of candidates specified in the Level Descriptors.*

Incongruent with assessment validity studies that relied mainly on quantitative evidence, this study adopted a mixed methodology, triangulating quantitative evidence with qualitative evidence through an analysis of the curriculum and assessment documents, live scripts, nominal group discussions among examiners and think-aloud protocols. A content analysis on the Assessment Objectives, assessment criteria of the examination papers and the Level Descriptors was conducted to justify the content validity (Validity Arguments (1)). Based on both secondary and primary sources, justification of Validity Arguments (2) and (3) was made. The secondary data sources comprised (i) a live script study, in which the scripts were re-scored in accordance with a Scoring Grid derived from the Level Descriptors of the examination; (ii) nominal group discussions and (iii) a retrospective think-aloud study as sources of evidence for the differentiation of performance and the sequential higher-order thinking process adopted by candidates. A more

comprehensive set of data covering all the Levels of Performance was gathered by re-scoring all the answers of the candidates in the live script study and scripts from the HKEAA Homepage by the same method deployed in the secondary data set. In the evaluation of the substantive validity, the alignment of the candidates' performance with Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the New Taxonomy (Marzano et al., 2008) and the KPI Model (Kuhn 2001, 2005) was analysed.

In Chapter 4, the validation process gathered evidence in support of the content alignment of the requirements of the 2015 LS Examination with the expected performance stipulated on Assessment Objectives and the Level Descriptors, though with an omission of the consideration of cultures/ values in the latter.

In Chapters 5 and 6, the substantive validity of the examination was evaluated with reference to the differentiation of candidate's performance on eight skill domains, namely *Understanding*, *Information-handling*, *Perspectives*, *Synthesis*, *Evaluation*, *Cultures/Values*, *Evidence* and *Communication*, as well as the sequence of higher-order thinking processes. The findings from the ANOVA analyses lent support to the differentiation of the candidates' performance on all skill domains between the 5 Levels, with the exception of *Evaluation* and *Cultures/Values* between the lowest 2 levels (Levels 1 and 2). In comparison with scores for other skills, the scores for these two exceptions were squeezed downwards, with the mean scores of candidates attaining Level 5 far below the maximum (5 points) on the scale (*Evaluation*: 3.667 points and *Cultures/Values*: 3.717 points) (Table 5.2) and the scores for Levels 1 and 2 being equivalent to the minimum (1 point).

The scaling down of *Evaluation* and *Cultures/Values* did not directly undermine the substantive

validity of the examination. In view of the use of the Level Descriptors in a holistic manner in the grading process, the overall scores (the mean of all the eight domains) of the live scripts were analysed statistically. Statistically significant differences were found in the overall scores among all the five Levels of Performance, providing evidence for the appropriateness of the Level Descriptors in differentiating the performance of candidates between the five levels.

For investigating the alignment between the assessment and cognitive models: Bloom's Taxonomy (1956), the Revised Taxonomy (Anderson & Krathwohl, 2001), the New Taxonomy (Marzano et al., 2008), correlations, analyses of constructed binary variables and a qualitative thematic analysis were conducted on the domains of *Knowledge*, *Information-handling*, *Synthesis* and *Evaluation*. Evidence for the alignment of the assessment with the cognitive complexity stipulated by the cognitive models was found. The differentiation of candidates' performance in accordance with a decreasing level of complexity from *Evaluation* to *Synthesis*, followed by *Information-handling* was supported by the quantitative and qualitative analysis. Among these three cognitive skills, *Information-handling* was the least demanding, while *Evaluation* the most demanding, aligning with the cognitive models. After dichotomising the scores, it was observed that the lower order skill, *Information-handling* was even mastered by less able candidates, Level 3 candidates. At the other end of the scale, candidates attaining higher Levels of Performance, Level 5, were able to master higher-order thinking skills, including the application of more complex *Principles* and *Generalisations of Knowledge*, *Synthesis* and *Evaluation*.

The thematic analysis showed that only the Level 5 candidate in the live script study was able to incorporate in his/her answers *Principles* and *Generalisations* from a wider perspective, including the personal, local, national and global perspectives, and in relation to a variety of social or economic cultures and values. In addition, a coherent *Synthesis* of high-level *Knowledge* and

Information-handling and sound rebuttals, as well as *Evaluation* in accordance with clearly delineated assessment criteria were only found in the answer scripts of the Level 5 candidate. In short, the examination awarded candidates who mastered higher-order thinking skills better with a higher Level of Performance. Evidence has been gathered to show that the examination distinguished candidates in accordance with the cognitive models, lending support to the substantive validity.

The thematic analysis of the think-aloud study complemented the live script study in contributing to the evidence in Chapter 6 for the sequential higher-order thinking processes demonstrated by Level 5 or above candidates. The candidates were able to integrate the findings from the data analysis and *Dispositions* of related *Knowledge* and *Values* in the formulation of evidence-based arguments in response to the questions, fulfilling the criteria for high-level reasoning as put forward by Kuhn (2001, 2005). The think-aloud study also supplemented the live script study with information about the highest level of cognitive skill in the New Taxonomy, *Meta-level thinking*. Level 5 or above candidates reviewed the accuracy of the analysis, the logicality and structure of the arguments to refine the quality of their answers.

As discussed in Chapters 7, this study put into practice an assessment validation process on the 2015 HKDSE LS Examination, by gathering evidence from multiple sources and from the perspectives of both content and substantive validity. Not only did the multiple sources of evidence complement one another, they also allowed triangulation of evidence, enhancing the validity of the process. Since stakeholders attach importance to high-stake examinations, the future careers and study paths of a large number of candidates hinge upon the examination results, the validity of these examinations has to be evaluated in a vigorous manner. The Argument-based Approach (Kane, 2013, 2015), which emphasises the gathering of evidence to justify the Validity

Argument, is practicable as illustrated in this study.

Validity evaluation does not only serve as a response to stakeholders' concerns, it can also inform test developers and teachers of directions for enhancement. From this study of the 2015 HKDSE LS Examination, areas that deserve attention were identified. Firstly, a description of the application of *Dispositions* of cultures and values was omitted in the Level Descriptors. To reflect the actual performance of candidates as stipulated in the Assessment Objectives, descriptions on the range of *Dispositions* of cultures and values should be specified by the Level Descriptors. Secondly, the domain of *Evidence* on the Level Descriptors should be rephrased in terms of Kuhn's (2001, 2005): "integration of evidence" at Level 5. Last but not least, teachers' attention should be drawn to the finding that the performance on *Evaluation* of candidates attaining Level 5 was below the expectations stipulated in the Level Descriptors.

For enhancing test development and students' learning, there is still much room for research on the assessment validation process. The methodological epistemology of assessment validation could be further explored to fill the gap in the literature. Research may go along this avenue to overcome the limitations of this study. The design of the retrospective think-aloud study could be enhanced to elicit evidence of the thinking processes adopted by candidates in answering examination questions. More details of the use of *Dispositions* and evidence in *Argument-formulation* may also be investigated by the think-aloud study. As Kuhn's (2001) research was centred around the use of evidence in *Argument-formulation* only, research on various types of arguments, for instance, the skill found to be the most demanding to candidates in this study, *Evaluation*, is worth investigating. Difficulties experienced by candidates to evaluate with concrete criteria might be further explored. Based on findings from these studies, test developers and educators may identify areas for further improvements in relation to the assessment design

and the weakness of candidates' thinking processes. Research along these lines may also contribute to the epistemology of assessment design.

In view of the difficulties in aligning the marking standards among examiners, nominal group discussions could be further analysed to illuminate the consensus-building process. This will not only enhance the quality of evidence from the re-scoring of live scripts in the proposed validation process, it can also add to the scarce literature on the methodological approaches to assessment validation based on qualitative evidence. Besides, this kind of research may contribute to the consensus-building among markers and examiners in the marking and grading processes of examinations, enhancing their validity.

The transferability of the assessment validation process adopted in this study can also be explored further. While most of the validation evaluation research studies in the literature are focusing on skill-based subjects, like English Language, there is much room for research in terms of the validity of large-scale high-stakes assessments. For instance, how can this validation process be adapted to evaluate the validity of content-based subjects, including Geography and Biology? Or how can the validity of the component of School-based Assessment in a high-stakes examination be evaluated?

This study has opened up an avenue of research on the practical side of validity evaluation on large-scale high-stake assessments, which has not been fully explored. As the "interpretation and use" of high-stake assessments exerts immense impact on the future careers of the candidates and are under the spotlight in society, further research on the validity evaluation practice for enhancing test development is valuable.

References

Alexander, P. A. et al. (2011). Higher order thinking and knowledge: domain-general and domain-specific trends and future directions. In G. J.Schraw &D. H.Robinson (Eds.), *Assessment of higher order thinking skills* (p. 418). Information Age Pub.

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., &Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, L. W., &Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives*. Longman.

Anderson, L. W., Sosniak, L. A., Bloom, B. S. (Benjamin S., &National Society for the Study of Education. (1994). *Bloom's taxonomy : a forty-year retrospective*. NSSE.

AQA. (2014a). *GCE AS and A Level Citizenship Studies Specification*. http://filestore.aqa.org.uk/subjects/specifications/alevel/AQA-2100-W-SP-14.PDF

AQA. (2014b). *GCE AS and A Level Specification General Studies A*. http://filestore.aqa.org.uk/subjects/specifications/alevel/AQA-2760-W-SP-14.PDF

Baird, J. A., Meadows, M., Leckie, G., &Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy and Practice*. https://doi.org/10.1080/0969594X.2015.1108283

Biddle, C., &Schafft, K. A. (2015). Axiology and Anomaly in the Practice of Mixed Methods Work: Pragmatism, Valuation, and the Transformative Paradigm. *Article Journal of Mixed Methods Research*, *9*(4), 320–334. https://doi.org/10.1177/1558689814533157

Biggs, J. B. (John B. (1982). *Evaluating the quality of learning : the SOLO taxonomy (structure of the observed learning outcome)*. Academic Press.

Bloom, B. S. (1956). *Taxonomy of educational objectives : the classification of educational goals. Handbook 1, Cognitive domain / by a committee of college and university examiners; Benjamin S. Bloom, editor. - 44BU* (Longman Group).

Bloom, B. S. (Benjamin S., Hastings, J. T. (John T., &Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple... *Language Testing*, *29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Charters, E. (2003). The Use of Think-aloud Methods in Qualitative Research: An Introduction to Think-aloud Methods. *Brock Education*, *12*(2), 68–82.

Cheung, C. K. (2009). Integrating media education into liberal studies: a positive response to curriculum reform in Hong Kong - HKU. *The Curriculum Journal*, *20*(4), 437–446.

Chiu, S. W., Yuen, K., &Leung, Y. (2018). How did Liberal Studies affect secondary school students. *Hong Kong and Macau Studies*, *2018*(1), 67–73. https://repository.eduhk.hk/en/publications/通識科如何影響香港中學生

Coniam, D. (2011). A qualitative examination of the attitudes of Liberal Studies markers towards onscreen marking in Hong Kong. *British Journal of Educational Technology*, *42*(6), 1042–1054. https://doi.org/10.1111/j.1467-8535.2010.01136.x

Coniam, D., &Falvey, P. (2016). *Validating technological innovation : the introduction and implementation of onscreen marking in Hong Kong*. Springer.

Coniam, D., &Yeung, S. A. (2010). Markers' perceptions regarding the onscreen marking of Liberal Studies in the Hong Kong public examination system. *Asia Pacific Journal of Education*, *30*(3), 249–271. https://doi.org/10.1080/02188791.2010.495836

Cook, D. A., Kuper, A., Hatala, R., &Ginsburg, S. (2016). When Assessment Data Are Words. *Academic Medicine*, *91*(10), 1359–1369. https://doi.org/10.1097/ACM.0000000000001175

Corliss, S. B., &Linn, M. C. (2011). Assessing Learning From Inquiry Science Instruction. In G. J.Schraw &D. H.Robinson (Eds.), *Assessment of higher order thinking skills* (p. 418). Information Age Pub.

Creswell, J. W. (2014). *Research design : qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Crisp, V., &Shaw, S. (2012). Applying Methods to Evaluate Construct Validity in the Context of A Level Assessment. *Educational Studies*, *38*(2), 209–222. https://doi.org/10.1080/03055698.2011.598670

Curriculum Development Council, Hong Kong Examinations and Assessment Authority, &Education Bureau. (2015). *Continual Renewal from Strength to Strength - Report on the New Academic Structure Medium-term Review and Beyond Curriculum Development Council*. https://334.edb.hkedcity.net/doc/eng/MTR_Report_e.pdf

DeVen, A. H.Van, Delbecq, A. L., Van DeVen, A. H., &Delbecq, A. L. (1974). The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes Decision Making Processes'. *Source: The Academy of Management Journal*, *17*(4), 605–621. http://www.jstor.org/stable/255641

DeLuca, C. (2011). Interpretive validity theory: mapping a methodology for validating educational assessments. *Educational Research*, *53*(3), 303–320. https://doi.org/10.1080/00131881.2011.598659

Deng, Z. (2009). The formation of a school subject and the nature of curriculum content: an analysis of liberal studies in Hong Kong. *Journal of Curriculum Studies*, *41*(5), 585–604.

Ebel, R. L. (1965). *Measuring Educational Achievement*. Prentice-Hall.

Ebel, R. L., &Frisbie, D. A. (1991). *Essentials of educational measurement*. Prentice Hall.

Feilzer, Y. M. (2010). Doing Mixed Methods Research Pragmatically: Implications for the Rediscovery of Pragmatism as a Research Paradigm. *Journal of Mixed Methods Research*, *4*(1), 6–16. https://doi.org/10.1177/1558689809349691

Flick, U. (2014). *An introduction to qualitative research* (5th ed., p. 587). SAGE.

Fung, D. (2016). Expectations versus reality: the case of Liberal Studies in Hong Kong's new senior secondary reforms. *Compare: A Journal of Comparative and International Education*, *46*(4), 624–644. https://doi.org/10.1080/03057925.2014.970009

Fung, D., &Howe, C. (2012). Liberal Studies in Hong Kong: A new perspective on critical thinking through group work. *Thinking Skills and Creativity*, *7*(2), 101–111. https://doi.org/10.1016/j.tsc.2012.04.002

Gardner, J. (2012). *Assessment and learning*. SAGE.

Goldstein, H. (2015). Validity, science and educational measurement. *Assessment in Education: Principles, Policy & Practice*, *22(2)*, 193–201. https://doi.org/10.1080/0969594X.2015.1015402

Hanushek, E. A. (2009). The economic value of education and cognitive skills. In Gary Sykes ; Barbara L Schneider ; David Nathan Plank 1954- ; Timothy G Ford (Ed.), *Handbook of education policy research* (pp. 39–56). Routledge.

Hauenstein, A. D. (1998). *A conceptual framework for educational objectives : a holistic approach to traditional taxonomies*. University Press of America.

Johnson, J., &Hayward, G. (2009). *Expert Group Report for Award Seeking Admission to the UCAS Tariff: Hong Kong Diploma of Secondary Education*.

Kane, M. T. (2013). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, *50*(1), 115–122. https://doi.org/10.1111/jedm.12007

Kane, M. T. (2015). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *January*, 1–14. https://doi.org/10.1080/0969594x.2015.1060192

Kerr, D., Keating, A., &Ireland, E. (2009). *Pupil assessment in citizenship education: purposes, practices and possibilities*. 117.

Kuhn, D. (2001). How do People Know? *Psychological Science*, *12*(1), 1–8. https://doi.org/10.1111/1467-9280.00302

Kuhn, D. (2005). *Education for thinking*. Harvard University Press. http://www.worldcat.org/title/education-for-thinking/oclc/718554515&referer=brief_results

Kuo, S. A. (2007). Which Rubric is More Suitable for NSS Liberal Studies ? Analytic or Holistic ? *Hong Kong Educational Research Association*, *22*(2), 179–199.

Lai, E., &Lam, C.-C. (2011). Learning to teach in a context of education reform: liberal studies student teachers' decision-making in lesson planning - HKU. *Journal of Education for Teaching*, *37*(2), 219–236.

Lane, S. (2005). Validity of High-Stakes Assessment: Are Students Engaged in Complex Thinking? *Educational Measurement: Issues and Practice*, *23*(3), 6–14. https://doi.org/10.1111/j.1745-3992.2004.tb00160.x

Legislative Council Secretariat. (2017). Legislative Council Panel on Education: Minutes of Meeting held on Monday, 13 February 2017. In *Legislative Council LC Paper No. CB* (Issue 4). https://www.legco.gov.hk/yr16-17/english/panels/ed/minutes/ed20170213.pdf

Leighton, J. P. (2011). A Cognitive Model for the Assessment of Higher Order Thinking. In Gregory J Schraw ; Daniel R Robinson (Ed.), *Assessment of higher order thinking skills* (pp. 151–181). Information Age Pub.

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.

Leighton, J. P., &Gierl, M. J. (2007). *Cognitive diagnostic assessment for education : theory and applications*. Cambridge University Press.

Leighton, J. P., &Gierl, M. J. (2011). *The learning sciences in educational assessment : the role of cognitive models*. Cambridge University Press.

Leung, A. W. L. (2009). Curriculum Integration for Liberal Studies in Hong Kong - HKU. *Curriculum and Teaching*, *24*(1), 75–98.

Leung, L. S. (2013). An inquiry of teachers' perception on the relationship between higher-order thinking nurturing and Liberal Studies public assessment in Hong Kong. *Hong Kong Teachers' Centre Journal*, *12*.

Leung, L. S. (2017). *Aligning summative assessment with curriculum aims of liberal studies in Hong Kong / Leung Lai Sim. - 44BU*.

Leung, T. Y. G. (2017). *Dissertation Proposal: An Evaluation of the Validity of the Hong Kong Diploma of Education Examination of Liberal Studies—a Case of a Large-scale Assessment*.

Lewin, C. (2011). Understanding and describing quantitative data. In C.Lewin &B.Somekh (Eds.), *Theory and methods in social research* (pp. 220–230). SAGE.

Lim, H. J. (2014). *Exploring the validity evidence of the TOEFL iBT reading test from a cognitive perspective*.

Linn, R. L., Gronlund, N. E., &Davis, K. M. (2000). *Measurement and assessment in teaching*. Merrill.

Mahon, E. A. (2006). High-Stakes Testing and English Language Learners: Questions of Validity. *Bilingual Research Journal*, *30*(2), 479–497. https://doi.org/10.1080/15235882.2006.10162886

Manns, M. et al. (2018). *The Culture of Testing Sociocultural Impacts on Learning in Asia and the Pacific*. http://www.unesco.

Marzano, R. J., Kendall, J. S., Administrators., A. A. of S., (U.S.), N. A. of E. S. P., &(U.S.), N. A. of S. S. P. (2008). Designing & assessing educational objectives: applying the new taxonomy. In *Designing and assessing educational objectives*. Corwin Press.

Mason, E. J. (2007). Measurement Issues in High Stakes Testing. *Journal of Applied School Psychology*, *23*(2), 27–46. https://doi.org/10.1300/J370v23n02_03

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, *50*(9), 741–749.

Mitchell, G. R., &Mitchell, G. R. (2003). Did Habermas Cede Nature to the Positivists? *Philosophy and Rhetoric*, *36*(1), 1–21. https://doi.org/10.1353/par.2003.0015

Morgan, D. L. (2007). Paradigms Lost and Pragmatism Regained. *Journal of Mixed Methods Research*. https://doi.org/10.1177/2345678906292462

Morris, P., &Chan, K. K. (1997). Cross-Curricular Themes and Curriculum Reform in Hong Kong: Policy as Discourse - HKU. *British Journal of Educational Studies*, *45*(3), 248–262.

Morse, J. M. (2000). Determining Sample Size. *Qualitative Health Research* , *10*(1), 3–5. http://journals.sagepub.com/doi/pdf/10.1177/104973200129118183

Moss, P. A., Girard, B. J., &Haniford, L. C. (2006). Validity in Educational Assessment. In *Special Issue on Rethinking Learning: What Counts as Learning and What Learning Counts* (Vol. 30). https://www-jstor-org.bris.idm.oclc.org/stable/pdf/4129771.pdf?refreqid=excelsior%3A8661d129ea459d6953be400baa369780

Newton, P. E., Shaw, S. D., &Cambridge Assessment. (2014). *Validity in educational & psychological assessment*. http://www.worldcat.org/title/validity-in-educational-psychological-assessment/oclc/881511346

Nielsen, J., Clemmensen, T., &Yssing, C. (2002). Getting access to what goes on in people's heads? - Reflections on the think-aloud technique. *NordiCHI*, *October*, *19-23*, 101–110.

Ormerod, R. (2006). The history and ideas of pragmatism. *Journal of the Operational Research Society*, *57*(8), 892–909. https://doi.org/10.1057/palgrave.jors.2602065

Pearson Edexcel. (2014). *Specification GCE General Studies Pearson Edexcel Level 3 Advanced Subsidiary GCE in General Studies (8GS01) Pearson Edexcel Level 3 Advanced GCE in General Studies (9GS01)*. https://qualifications.pearson.com/content/dam/pdf/A Level/General Studies/2013/Specification and sample assessments/UA035233_GCE_Lin_GenStu_Issue_3.pdf

Pellegrino, J. W., &Wilson, M. (2015). Assessment of Complex Cognition: Commentary on the Design and Validation of Assessments. *Theory Into Practice*, *5841*(July), 0–0. https://doi.org/10.1080/00405841.2015.1044377

Pellegrino, James W. (2016). Introduction to Special Section of *Educational Psychologist* on Educational Assessment: Validity Arguments and Evidence—Blending Cognitive, Instructional, and Measurement Models and Methods. *Educational Psychologist*, *51*(1), 57–58. https://doi.org/10.1080/00461520.2016.1150786

Pellegrino, James W., Chudowsky, N., Glaser, R., &National Research Council (U.S.). Committee on the Foundations of Assessment. (2001). *Knowing what students know : the science and design of educational assessment*. National Academy Press.

Pellegrino, James W., DiBello, L.V., &Goldman, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist*, *1520*(April), 1–23. https://doi.org/10.1080/00461520.2016.1145550

Posner, G. J., Strike, K. A., Hewson, P. W., &Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227. https://doi.org/10.1002/sce.3730660207

Punch, K. (2014). *Introduction to social research : quantitative & qualitative approaches*.

Rhoten, D., Mansilla, V. B., Chun, M., &Klein, J. T. (2000). *Interdisciplinary Education at Liberal Arts Institutions Teagle Foundation White Paper*.

Robson, C. (2011). *Real world research : a resource for users of social research methods in applied settings* (Wiley). Wiley.

Schilling, K. L. (1987). *Assessing models of Liberal Education: an Empirical Comparison*.

Schraw, G. J., &Robinson, D. H. (2011). *Assessment of higher order thinking skills*. Information Age Pub.

Shannon-Baker, P. (2016). Making Paradigms Meaningful in Mixed Methods Research. *Article Journal of Mixed Methods Research*, *10*(4), 319–334. https://doi.org/10.1177/1558689815575861

Shaw, S., Crisp, V., &Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, *19*(2), 159–176. https://doi.org/10.1080/0969594X.2011.563356

Shaw, S., &Imam, H. (2013a). Assessment of international students through the medium of english: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly*, *10*(4), 452–475. https://doi.org/10.1080/15434303.2013.866117

Shaw, S., &Imam, H. (2013b). Assessment of International Students Through the Medium of English: Ensuring Validity and Fairness in Content-Based Examinations. *Language Assessment Quarterly*, *10*(4), 452–475.

Someren, M. W. van., Barnard, Y. F., &Sandberg, J. (Jacobijn). (1994). *The think aloud method : a practical guide to modelling cognitive processes*. Academic Press.

Stobaugh, R. (2014). *Assessing critical thinking in middle and high schools : meeting the common core*. Eye on Education.

The Curriculum Development Council and the Hong Kong Examinations and Assessment Authority. (2014). *Liberal Studies Curriculum and Assessment Guide*. http://334.edb.hkedcity.net/doc/eng/curriculum/LS C&A Guide_updated_e.pdf

The Hong Kong Examinations and Assessment Authority. (2007). *Development of the Draft Level Descriptors for HKSDE LS*.

The Hong Kong Examinations and Assessment Authority. (2014). *Level Descriptors (Revised)*. http://www.hkeaa.edu.hk/en/hkdse/assessment/subject_information/category_a_subjects/hkdse_subj.html?A1&1&3_4

The Hong Kong Examinations and Assessment Authority. (2015a). *HKDSE - Liberal Studies Examination Report and Question Papers (with Marking Schemes)*. The Hong Kong Examinations and Assessment Authority.

The Hong Kong Examinations and Assessment Authority. (2015b). *Samples of Candidates' Performance*. http://www.hkeaa.edu.hk/en/HKDSE/assessment/subject_information/category_a_subjects/lib_st/sp/2015.html

The Hong Kong Examinations and Assessment Authority. (2017). *Assessment Framework (Liberal Studies)*. http://www.hkeaa.edu.hk/DocLibrary/HKDSE/Subject_Information/lib_st/2017hkdse-e-ls.pdf

The Hongkong Federation of Youth Groups. (2018). *Improvements on the Teaching and Learning of Liberal Studies in Senior Secondary Education*.

Trainor, A. A., &Graue, E. (2014). Evaluating rigor in qualitative methodology and research dissemination. *Remedial and Special Education*. https://doi.org/10.1177/0741932514528100

Usher, R. (1996). A Critique of the Neglected Epistemological Assumptions of Educational Research. In D.Scott &R.Usher (Eds.), *Understanding Educational Research*. Routledge.

Webb, N. L. (2002). *Depth-of-Knowledge Levels for Four Content Areas*.

Wyse, Adam E., Viger, S. G. (2011). How Item Writers Understand Depth of Knowledge. *Educational Assessment*, *16*(4), 185–206. https://doi.org/10.1080/10627197.2011.634286

Zahedi, K., Shamsaee, S., Zahedi, K., &Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): evidence from Iranian test-takers. *Educ Asse Eval Acc*, *24*, 263–277. https://doi.org/10.1007/s11092-011-9137-z

Appendix I

Live Script Study (2015) – Scoring Grid

Please tick ONE box only for each row to indicate the best description of the skills demonstrated in the script.

Domain of Skill	Description of the mastery of the skill				
Understanding and application of relevant knowledge, key ideas and concepts of the subject	comprehensive knowledge, understanding of key ideas and concepts	broad knowledge, understanding of key ideas and concepts	general knowledge, understanding of key ideas and concepts	basic knowledge, understanding of key ideas and concepts	elementary knowledge, understanding of key ideas and concepts
	1A	1B	1C	1D	1E
Handling of relevant information	generalise information*	analyse information**	interpret information	identify relevant information	Identify some basic and simple information
	2A	2B	2C	2D	2E
Interpretation and analysis of the interdependence among personal, local, national and global issues	interpret and analyse coherently from different perspectives	interpret and analyse from different perspectives	interpret appropriately from different perspectives	interpret briefly from some perspectives	identify simple relationships from a few perspectives
	3A	3B	3C	3D	3E

Appendix I

Domain of Skill	Description of the mastery of the skill				
Formulation of viewpoints, opinions and suggestions	synthesise their own opinions/ suggestions on the basis of logical arguments	synthesise their own opinions/ suggestions with partly reasonable arguments	elaborate on opinions/ suggestions from the sources with partly reasonable arguments/ with simple elaboration	give irrelevant opinions, suggestions and ungrounded arguments	
	4A	4B	4C	4D	
	evaluate various viewpoints/ entities or assess impacts/ effectiveness/ relationships based on clear criteria/ standards	compare viewpoints/ entities or explain the impacts/ effectiveness/ relationships without clear criteria/ standards	explain various viewpoints/ entities separately or explain impacts/ effectiveness/ relationships by simple arguments	provide one-sided arguments/ describe one of the entities/ pros/cons without comparison/ evaluation/ assessment	
	5A	5B	5C	5D	
	show appreciation of different cultures/ universal values; or shows empathy/ open-mindedness/ tolerance towards a wide range of people/ incidents/ views / values in the formulation of arguments	consider particular cultures/ universal values; or show empathy/ open-mindedness/ tolerance towards particular groups of people/ types of incidents/ views/ values in the formulation of arguments	show limited awareness of different cultures/ universal values, the concerns/ situations of different groups of people in the formulation of arguments	elaborate on their own views based on their own/ values/ cultures; without sound justification	
	6A	6B	6C	6D	
Respect for evidence	conceptualise evidence or use sufficient examples	identify some evidence or use some examples	identify limited evidence or use a few examples	use irrelevant examples/ without examples or give little/ no evidence	
	7A	7B	7C	7D	
Communication of ideas	communicate concisely, logically and systematically	communicate logically and systematically	communicate in an organized manner	communicate simple ideas	Express simple ideas briefly
	8A	8B	8C	8D	8E

Notes:

*Generalising information: looking for patterns or connection in the information

**Analysing information may include: breaking material into constituent parts; identifying similarities and differences, categorising; identify conclusion and supporting statements

1. The complete ANOVA output on the scores of skill domains by level

1A. ANOVA statistics of scores of skill domains of scripts from the joint study (Levels 3 to 5)

Table AII-1A-1: Descriptives of scores of skill domains of scripts from the joint study (Levels 3 to 5)

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
1 Understanding	3	40	3.28750	.703813	.111283	3.06241	3.51259	1.000	4.500
	4	112	3.76116	.607191	.057374	3.64747	3.87485	1.000	5.000
	5	136	4.45864	.534256	.045812	4.36804	4.54924	3.000	5.000
	Total	288	4.02474	.732286	.043150	3.93981	4.10967	1.000	5.000
2 Information-handling	3	30	3.73333	.712975	.130171	3.46710	3.99956	2.000	5.000
	4	84	4.19643	.634205	.069197	4.05880	4.33406	3.000	5.000
	5	102	4.71078	.430497	.042626	4.62623	4.79534	3.500	5.000
	Total	216	4.37500	.659016	.044840	4.28662	4.46338	2.000	5.000
3 Perspectives	3	40	3.15625	.722104	.114175	2.92531	3.38719	1.000	4.000
	4	112	3.80804	.601221	.056810	3.69546	3.92061	1.000	5.000
	5	136	4.42279	.518647	.044474	4.33484	4.51075	3.000	5.000
	Total	288	4.00781	.731734	.043118	3.92295	4.09268	1.000	5.000
4 Synthesis	3	40	3.04938	.690561	.109187	2.82852	3.27023	1.975	4.325
	4	112	3.48683	.562291	.053131	3.38155	3.59211	2.000	5.000
	5	136	4.25110	.646874	.055469	4.14140	4.36080	2.550	5.000
	Total	288	3.78698	.772641	.045528	3.69737	3.87659	1.975	5.000
5 Evaluation	3	39	2.51731	.763974	.122334	2.26966	2.76496	1.000	3.600
	4	111	2.94414	.714787	.067845	2.80969	3.07860	1.000	5.000

Appendix II

	5	136	3.66691	.704050	.060372	3.54751	3.78631	2.300	5.000
	Total	286	3.22963	.837998	.049552	3.13210	3.32717	1.000	5.000
6 Cultures/Values	3	36	2.80972	.717717	.119619	2.56688	3.05256	1.000	3.900
	4	109	2.99610	.669881	.064163	2.86892	3.12328	1.000	4.300
	5	135	3.71676	.680830	.058596	3.60087	3.83265	2.225	5.000
	Total	280	3.31960	.782117	.046740	3.22759	3.41161	1.000	5.000
7 Evidence	3	40	3.09688	.778974	.123167	2.84775	3.34600	1.000	4.575
	4	112	3.36763	.626720	.059219	3.25029	3.48498	2.300	5.000
	5	136	4.25708	.626026	.053681	4.15091	4.36324	2.300	5.000
	Total	288	3.75004	.810597	.047765	3.65603	3.84406	1.000	5.000
8 Communication	3	40	3.50000	.609750	.096410	3.30499	3.69501	2.000	4.750
	4	112	3.97321	.460558	.043519	3.88698	4.05945	2.000	5.000
	5	136	4.60993	.428512	.036745	4.53726	4.68260	3.500	5.000
	Total	288	4.20816	.621979	.036650	4.13602	4.28030	2.000	5.000
Overall Average	3	40	3.12177	.575005	.090916	2.93787	3.30566	1.767	4.184
	4	112	3.55138	.465100	.043948	3.46429	3.63846	1.757	4.825
	5	136	4.24190	.435669	.037358	4.16802	4.31579	2.850	5.000
	Total	288	3.81779	.631130	.037190	3.74459	3.89099	1.757	5.000

Table AII-1A-2: ANOVA statistics of scores of skill domains of scripts from the joint study (Levels 3 to 5)

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
1 Understanding	Between Groups	55.127	2	27.563	79.529	.000
	Within Groups	98.775	285	.347		
	Total	153.902	287			
2 Information-handling	Between Groups	26.531	2	13.266	42.271	.000
	Within Groups	66.844	213	.314		
	Total	93.375	215			
3 Perspectives	Between Groups	56.897	2	28.448	83.782	.000
	Within Groups	96.773	285	.340		
	Total	153.670	287			
4 Synthesis	Between Groups	61.148	2	30.574	79.083	.000
	Within Groups	110.183	285	.387		
	Total	171.331	287			
5 Evaluation	Between Groups	54.841	2	27.420	53.407	.000
	Within Groups	145.298	283	.513		
	Total	200.139	285			
6 Cultures/Values	Between Groups	42.060	2	21.030	45.296	.000
	Within Groups	128.606	277	.464		
	Total	170.666	279			
7 Evidence	Between Groups	68.407	2	34.204	81.118	.000
	Within Groups	120.171	285	.422		
	Total	188.578	287			
8 Communication	Between Groups	48.195	2	24.097	109.300	.000

	Within Groups	62.834	285	.220		
	Total	111.028	287			
Overall Average	Between Groups	51.790	2	25.895	118.024	.000
	Within Groups	62.530	285	.219		
	Total	114.319	287			

Table AII-1A-3: Post Hoc Tests of scores of skill domains of scripts from the joint study (Levels 3 to 5)

Post Hoc Tests

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) Level	(J) Level	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1 Understanding	3	4	-.473661*	.108439	.000	-.72915	-.21818
		5	-1.171140*	.105891	.000	-1.42062	-.92166
	4	3	.473661*	.108439	.000	.21818	.72915
		5	-.697479*	.075119	.000	-.87446	-.52050
	5	3	1.171140*	.105891	.000	.92166	1.42062
		4	.697479*	.075119	.000	.52050	.87446
2 Information-handling	3	4	-.463095*	.119150	.000	-.74431	-.18188
		5	-.977451*	.116350	.000	-1.25206	-.70284
	4	3	.463095*	.119150	.000	.18188	.74431
		5	-.514356*	.082539	.000	-.70916	-.31955
	5	3	.977451*	.116350	.000	.70284	1.25206
		4	.514356*	.082539	.000	.31955	.70916
3 Perspectives	3	4	-.651786*	.107334	.000	-.90467	-.39890
		5	-1.266544*	.104812	.000	-1.51348	-1.01960
	4	3	.651786*	.107334	.000	.39890	.90467
		5	-.614758*	.074354	.000	-.78994	-.43958
	5	3	1.266544*	.104812	.000	1.01960	1.51348
		4	.614758*	.074354	.000	.43958	.78994
4 Synthesis	3	4	-.437455*	.114530	.000	-.70729	-.16762

Appendix II

		5	-1.201728*	.111839	.000	-1.46522	-.93823
	4	3	.437455*	.114530	.000	.16762	.70729
		5	-.764273*	.079338	.000	-.95120	-.57735
	5	3	1.201728*	.111839	.000	.93823	1.46522
		4	.764273*	.079338	.000	.57735	.95120
5 Evaluation	3	4	-.426836*	.133379	.004	-.74109	-.11258
		5	-1.149604*	.130153	.000	-1.45626	-.84295
	4	3	.426836*	.133379	.004	.11258	.74109
		5	-.722768*	.091655	.000	-.93872	-.50682
	5	3	1.149604*	.130153	.000	.84295	1.45626
		4	.722768*	.091655	.000	.50682	.93872
6 Cultures/Vlues	3	4	-.186379	.130982	.330	-.49502	.12226
		5	-.907037*	.127812	.000	-1.20821	-.60586
	4	3	.186379	.130982	.330	-.12226	.49502
		5	-.720658*	.087742	.000	-.92741	-.51391
	5	3	.907037*	.127812	.000	.60586	1.20821
		4	.720658*	.087742	.000	.51391	.92741
7 Evidence	3	4	-.270759	.119608	.063	-.55256	.01104
		5	-1.160202*	.116798	.000	-1.43538	-.88502
	4	3	.270759	.119608	.063	-.01104	.55256
		5	-.889443*	.082856	.000	-1.08465	-.69423
	5	3	1.160202*	.116798	.000	.88502	1.43538
		4	.889443*	.082856	.000	.69423	1.08465
8 Communication	3	4	-.473214*	.086488	.000	-.67698	-.26945
		5	-1.109926*	.084456	.000	-1.30891	-.91095
	4	3	.473214*	.086488	.000	.26945	.67698

Appendix II

		5	-.636712*	.059913	.000	-.77787	-.49556
	5	3	1.109926*	.084456	.000	.91095	1.30891
		4	.636712*	.059913	.000	.49556	.77787
Overall Average	3	4	-.429612*	.086279	.000	-.63289	-.22634
		5	-1.120135*	.084252	.000	-1.31863	-.92164
	4	3	.429612*	.086279	.000	.22634	.63289
		5	-.690523*	.059768	.000	-.83134	-.54971
	5	3	1.120135*	.084252	.000	.92164	1.31863
		4	.690523*	.059768	.000	.54971	.83134

*. The mean difference is significant at the 0.05 level.

1B. ANOVA statistics of scores of skill domains by level (Levels 1 to 5)

Table AII-1B-1: Descriptives of scores of skill domains by level (Levels 1 to 5)

Descriptives

						95% Confidence Interval for Mean			
		N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
1 Understanding	1	24	1.00000	.000000	.000000	1.00000	1.00000	1.000	1.000
	2	24	2.00000	.000000	.000000	2.00000	2.00000	2.000	2.000
	3	64	3.17969	.571250	.071406	3.03699	3.32238	1.000	4.500
	4	112	3.76116	.607191	.057374	3.64747	3.87485	1.000	5.000
	5	136	4.45864	.534256	.045812	4.36804	4.54924	3.000	5.000
	Total	360	3.61979	1.104608	.058218	3.50530	3.73428	1.000	5.000
2 Information-handling	1	12	1.91667	.900337	.259905	1.34462	2.48871	1.000	3.000
	2	12	2.83333	.577350	.166667	2.46650	3.20016	2.000	4.000
	3	42	3.69048	.659845	.101816	3.48485	3.89610	2.000	5.000
	4	84	4.19643	.634205	.069197	4.05880	4.33406	3.000	5.000
	5	102	4.71078	.430497	.042626	4.62623	4.79534	3.500	5.000
	Total	252	4.14683	.902227	.056835	4.03489	4.25876	1.000	5.000
3 Perspectives	1	24	1.00000	.000000	.000000	1.00000	1.00000	1.000	1.000
	2	24	2.00000	.000000	.000000	2.00000	2.00000	2.000	2.000
	3	64	3.12891	.594272	.074284	2.98046	3.27735	1.000	4.000

	4	112	3.80804	.601221	.056810	3.69546	3.92061	1.000	5.000
	5	136	4.42279	.518647	.044474	4.33484	4.51075	3.000	5.000
	Total	360	3.61181	1.098793	.057911	3.49792	3.72569	1.000	5.000
4 Synthesis	1	24	1.00000	.000000	.000000	1.00000	1.00000	1.000	1.000
	2	24	2.30000	.000000	.000000	2.30000	2.30000	2.300	2.300
	3	64	3.09336	.664494	.083062	2.92737	3.25935	1.975	4.325
	4	112	3.48683	.562291	.053131	3.38155	3.59211	2.000	5.000
	5	136	4.25110	.646874	.055469	4.14140	4.36080	2.550	5.000
	Total	360	3.46069	1.043262	.054985	3.35256	3.56883	1.000	5.000
5 Evaluation	1	24	1.00000	.000000	.000000	1.00000	1.00000	1.000	1.000
	2	24	1.37917	.603597	.123209	1.12429	1.63404	1.000	2.300
	3	63	2.55833	.695101	.087574	2.38327	2.73339	1.000	3.600
	4	111	2.94414	.714787	.067845	2.80969	3.07860	1.000	5.000
	5	136	3.66691	.704050	.060372	3.54751	3.78631	2.300	5.000
	Total	358	2.91557	1.046640	.055317	2.80679	3.02436	1.000	5.000
6 Cultures/Values	1	21	1.00000	.000000	.000000	1.00000	1.00000	1.000	1.000
	2	22	1.29545	.557612	.118883	1.04822	1.54269	1.000	2.300
	3	58	2.54914	.827090	.108602	2.33167	2.76661	1.000	3.900
	4	109	2.99610	.669881	.064163	2.86892	3.12328	1.000	4.300
	5	135	3.71676	.680830	.058596	3.60087	3.83265	2.225	5.000
	Total	345	2.97301	1.057799	.056950	2.86099	3.08502	1.000	5.000

7 Evidence	1	24	1.00000	.000000	.000000	1.00000	1.00000	1.000	1.000
	2	24	1.75833	.654693	.133639	1.48188	2.03479	1.000	2.300
	3	64	2.89961	.841820	.105227	2.68933	3.10989	1.000	4.575
	4	112	3.36763	.626720	.059219	3.25029	3.48498	2.300	5.000
	5	136	4.25708	.626026	.053681	4.15091	4.36324	2.300	5.000
	Total	360	3.35531	1.143296	.060257	3.23681	3.47381	1.000	5.000
8 Communication	1	24	1.00000	.000000	.000000	1.00000	1.00000	1.000	1.000
	2	24	2.00000	.000000	.000000	2.00000	2.00000	2.000	2.000
	3	63	3.31746	.541068	.068168	3.18119	3.45373	2.000	4.750
	4	112	3.97321	.460558	.043519	3.88698	4.05945	2.000	5.000
	5	136	4.60993	.428512	.036745	4.53726	4.68260	3.500	5.000
	Total	359	3.76866	1.107737	.058464	3.65369	3.88364	1.000	5.000
Overall Average	1	24	1.05878	.100669	.020549	1.01627	1.10129	1.000	1.286
	2	24	1.89427	.198504	.040519	1.81045	1.97809	1.614	2.271
	3	64	3.02217	.494529	.061816	2.89864	3.14570	1.767	4.184
	4	112	3.55138	.465100	.043948	3.46429	3.63846	1.757	4.825
	5	136	4.24190	.435669	.037358	4.16802	4.31579	2.850	5.000
	Total	360	3.44151	1.000871	.052751	3.33778	3.54525	1.000	5.000

Table AII-1B-2: ANOVA statistics of scores of skill domains (Levels 1 to 5)

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
1 Understanding	Between Groups	338.022	4	84.505	299.949	.000
	Within Groups	100.015	355	.282		
	Total	438.037	359			
2 Information-handling	Between Groups	121.781	4	30.445	91.111	.000
	Within Groups	82.537	247	.334		
	Total	204.317	251			
3 Perspectives	Between Groups	334.751	4	83.688	301.047	.000
	Within Groups	98.686	355	.278		
	Total	433.437	359			
4 Synthesis	Between Groups	271.331	4	67.833	201.676	.000
	Within Groups	119.403	355	.336		
	Total	390.734	359			
5 Evaluation	Between Groups	229.623	4	57.406	125.510	.000
	Within Groups	161.455	353	.457		
	Total	391.078	357			
6 Cultures/Values	Between Groups	228.816	4	57.204	124.597	.000
	Within Groups	156.099	340	.459		

	Total	384.915	344			
7 Evidence	Between Groups	318.248	4	79.562	187.037	.000
	Within Groups	151.010	355	.425		
	Total	469.258	359			
8 Communication	Between Groups	372.810	4	93.203	496.262	.000
	Within Groups	66.485	354	.188		
	Total	439.295	358			
Overall Average	Between Groups	293.444	4	73.361	393.509	.000
	Within Groups	66.182	355	.186		
	Total	359.626	359			

Table AII-1B-3: Statistics of Post Hoc Tests of scores of skill domains (Levels 1 to 5)

Post Hoc Tests

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) Level	(J) Level	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
1 Understanding	1	2	-1.000000*	.153224	.000	-1.42012	-.57988
		3	-2.179688*	.127047	.000	-2.52803	-1.83134
		4	-2.761161*	.119392	.000	-3.08851	-2.43381
		5	-3.458640*	.117518	.000	-3.78086	-3.13642
	2	1	1.000000*	.153224	.000	.57988	1.42012
		3	-1.179688*	.127047	.000	-1.52803	-.83134
		4	-1.761161*	.119392	.000	-2.08851	-1.43381
		5	-2.458640*	.117518	.000	-2.78086	-2.13642
	3	1	2.179688*	.127047	.000	1.83134	2.52803
		2	1.179688*	.127047	.000	.83134	1.52803
		4	-.581473*	.083172	.000	-.80952	-.35343
		5	-1.278952*	.080459	.000	-1.49956	-1.05835
	4	1	2.761161*	.119392	.000	2.43381	3.08851
		2	1.761161*	.119392	.000	1.43381	2.08851

		3	.581473*	.083172	.000	.35343	.80952
		5	-.697479*	.067728	.000	-.88318	-.51178
	5	1	3.458640*	.117518	.000	3.13642	3.78086
		2	2.458640*	.117518	.000	2.13642	2.78086
		3	1.278952*	.080459	.000	1.05835	1.49956
		4	.697479*	.067728	.000	.51178	.88318
2 Information-handling	1	2	-.916667*	.235993	.001	-1.56518	-.26815
		3	-1.773810*	.189215	.000	-2.29378	-1.25384
		4	-2.279762*	.178394	.000	-2.76999	-1.78953
		5	-2.794118*	.176415	.000	-3.27891	-2.30932
	2	1	.916667*	.235993	.001	.26815	1.56518
		3	-.857143*	.189215	.000	-1.37711	-.33717
		4	-1.363095*	.178394	.000	-1.85333	-.87286
		5	-1.877451*	.176415	.000	-2.36225	-1.39266
	3	1	1.773810*	.189215	.000	1.25384	2.29378
		2	.857143*	.189215	.000	.33717	1.37711
		4	-.505952*	.109244	.000	-.80616	-.20575
		5	-1.020308*	.105982	.000	-1.31155	-.72907
	4	1	2.279762*	.178394	.000	1.78953	2.76999
		2	1.363095*	.178394	.000	.87286	1.85333
		3	.505952*	.109244	.000	.20575	.80616

		5	-.514356*	.085171	.000	-.74841	-.28030
	5	1	2.794118*	.176415	.000	2.30932	3.27891
		2	1.877451*	.176415	.000	1.39266	2.36225
		3	1.020308*	.105982	.000	.72907	1.31155
		4	.514356*	.085171	.000	.28030	.74841
3 Perspectives	1	2	-1.000000*	.152203	.000	-1.41732	-.58268
		3	-2.128906*	.126200	.000	-2.47493	-1.78288
		4	-2.808036*	.118596	.000	-3.13321	-2.48286
		5	-3.422794*	.116734	.000	-3.74286	-3.10273
	2	1	1.000000*	.152203	.000	.58268	1.41732
		3	-1.128906*	.126200	.000	-1.47493	-.78288
		4	-1.808036*	.118596	.000	-2.13321	-1.48286
		5	-2.422794*	.116734	.000	-2.74286	-2.10273
	3	1	2.128906*	.126200	.000	1.78288	2.47493
		2	1.128906*	.126200	.000	.78288	1.47493
		4	-.679129*	.082617	.000	-.90565	-.45261
		5	-1.293888*	.079923	.000	-1.51302	-1.07475
	4	1	2.808036*	.118596	.000	2.48286	3.13321
		2	1.808036*	.118596	.000	1.48286	2.13321
		3	.679129*	.082617	.000	.45261	.90565
		5	-.614758*	.067276	.000	-.79922	-.43030

	5	1	3.422794*	.116734	.000	3.10273	3. 74286
		2	2.422794*	.116734	.000	2.10273	2.74286
		3	1.293888*	.079923	.000	1.07475	1.51302
		4	.614758*	.067276	.000	.43030	.79922
4 Synthesis	1	2	-1.300000*	.167418	.000	-1.75904	-.84096
		3	-2.093359*	.138816	.000	-2.47397	-1.71275
		4	-2.486830*	.130451	.000	-2.84451	-2.12915
		5	-3.251103*	.128404	.000	-3.60317	-2.89904
	2	1	1.300000*	.167418	.000	.84096	1.75904
		3	-.793359*	.138816	.000	-1.17397	-.41275
		4	-1.186830*	.130451	.000	-1.54451	-.82915
		5	-1.951103*	.128404	.000	-2.30317	-1.59904
	3	1	2.093359*	.138816	.000	1.71275	2.47397
		2	.793359*	.138816	.000	.41275	1.17397
		4	-.393471*	.090876	.000	-.64264	-.14430
		5	-1.157744*	.087912	.000	-1.39879	-.91670
	4	1	2.486830*	.130451	.000	2.12915	2.84451
		2	1.186830*	.130451	.000	.82915	1.54451
		3	.393471*	.090876	.000	.14430	.64264
		5	-.764273*	.074002	.000	-.96717	-.56137
	5	1	3.251103*	.128404	.000	2.89904	3.60317

		2	1.951103*	.128404	.000	1.59904	2. 30317
		3	1.157744*	.087912	.000	.91670	1.39879
		4	.764273*	.074002	.000	.56137	.96717
5 Evaluation	1	2	-.379167	.195230	.297	-.91447	.15614
		3	-1.558333*	.162226	.000	-2.00315	-1.11352
		4	-1.944144*	.152243	.000	-2.36158	-1.52670
		5	-2.666912*	.149735	.000	-3.07747	-2.25635
	2	1	.379167	.195230	.297	-.15614	.91447
		3	-1.179167*	.162226	.000	-1.62398	-.73435
		4	-1.564977*	.152243	.000	-1.98242	-1.14754
		5	-2.287745*	.149735	.000	-2.69831	-1.87718
	3	1	1.558333*	.162226	.000	1.11352	2.00315
		2	1.179167*	.162226	.000	.73435	1.62398
		4	-.385811*	.106679	.003	-.67832	-.09330
		5	-1.108578*	.103068	.000	-1.39118	-.82597
	4	1	1.944144*	.152243	.000	1.52670	2.36158
		2	1.564977*	.152243	.000	1.14754	1.98242
		3	.385811*	.106679	.003	.09330	.67832
		5	-.722768*	.086508	.000	-.95997	-.48557
	5	1	2.666912*	.149735	.000	2.25635	3.07747
		2	2.287745*	.149735	.000	1.87718	2.69831

		3	1.108578*	.103068	.000	.82597	1.39118
		4	.722768*	.086508	.000	.48557	.95997
6 Cultures/Values	1	2	-.295455	.206716	.609	-.86237	.27146
		3	-1.549138*	.172564	.000	-2.02239	-1.07589
		4	-1.996101*	.161476	.000	-2.43895	-1.55326
		5	-2.716759*	.158945	.000	-3.15266	-2.28086
	2	1	.295455	.206716	.609	-.27146	.86237
		3	-1.253683*	.169660	.000	-1.71897	-.78839
		4	-1.700646*	.158369	.000	-2.13497	-1.26632
		5	-2.421305*	.155787	.000	-2.84855	-1.99406
	3	1	1.549138*	.172564	.000	1.07589	2.02239
		2	1.253683*	.169660	.000	.78839	1.71897
		4	-.446963*	.110126	.001	-.74898	-.14494
		5	-1.167621*	.106380	.000	-1.45936	-.87588
	4	1	1.996101*	.161476	.000	1.55326	2.43895
		2	1.700646*	.158369	.000	1.26632	2.13497
		3	.446963*	.110126	.001	.14494	.74898
		5	-.720658*	.087252	.000	-.95994	-.48137
	5	1	2.716759*	.158945	.000	2.28086	3.15266
		2	2.421305*	.155787	.000	1.99406	2.84855
		3	1.167621*	.106380	.000	.87588	1.45936

		4	.720658*	.087252	.000	.48137	.95994
7 Evidence	1	2	-.758333*	.188277	.001	-1.27456	-.24211
		3	-1.899609*	.156111	.000	-2.32764	-1.47158
		4	-2.367634*	.146705	.000	-2.76988	-1.96539
		5	-3.257077*	.144402	.000	-3.65301	-2.86115
	2	1	.758333*	.188277	.001	.24211	1.27456
		3	-1.141276*	.156111	.000	-1.56931	-.71324
		4	-1.609301*	.146705	.000	-2.01154	-1.20706
		5	-2.498744*	.144402	.000	-2.89467	-2.10281
	3	1	1.899609*	.156111	.000	1.47158	2.32764
		2	1.141276*	.156111	.000	.71324	1.56931
		4	-.468025*	.102199	.000	-.74824	-.18781
		5	-1.357468*	.098865	.000	-1.62854	-1.08639
	4	1	2.367634*	.146705	.000	1.96539	2.76988
		2	1.609301*	.146705	.000	1.20706	2.01154
		3	.468025*	.102199	.000	.18781	.74824
		5	-.889443*	.083222	.000	-1.11762	-.66126
	5	1	3.257077*	.144402	.000	2.86115	3.65301
		2	2.498744*	.144402	.000	2.10281	2.89467
		3	1.357468*	.098865	.000	1.08639	1.62854
		4	.889443*	.083222	.000	.66126	1.11762

8 Communication	1	2	-1.000000*	.125103	.000	-1.34302	-.65698
		3	-2.317460*	.103954	.000	-2.60249	-2.03243
		4	-2.973214*	.097480	.000	-3.24049	-2.70594
		5	-3.609926*	.095950	.000	-3.87301	-3.34684
	2	1	1.000000*	.125103	.000	.65698	1.34302
		3	-1.317460*	.103954	.000	-1.60249	-1.03243
		4	-1.973214*	.097480	.000	-2.24049	-1.70594
		5	-2.609926*	.095950	.000	-2.87301	-2.34684
	3	1	2.317460*	.103954	.000	2.03243	2.60249
		2	1.317460*	.103954	.000	1.03243	1.60249
		4	-.655754*	.068249	.000	-.84289	-.46862
		5	-1.292466*	.066046	.000	-1.47356	-1.11138
	4	1	2.973214*	.097480	.000	2.70594	3.24049
		2	1.973214*	.097480	.000	1.70594	2.24049
		3	.655754*	.068249	.000	.46862	.84289
		5	-.636712*	.055298	.000	-.78833	-.48509
	5	1	3.609926*	.095950	.000	3.34684	3.87301
		2	2.609926*	.095950	.000	2.34684	2.87301
		3	1.292466*	.066046	.000	1.11138	1.47356
		4	.636712*	.055298	.000	.48509	.78833
Overall Average	1	2	-.835491*	.124642	.000	-1.17724	-.49374

		3	-1.963391*	.103348	.000	-2.24675	-1. 68003
		4	-2.492599*	.097120	.000	-2.75889	-2.22631
		5	-3.183122*	.095596	.000	-3.44523	-2.92101
	2	1	.835491*	.124642	.000	.49374	1.17724
		3	-1.127899*	.103348	.000	-1.41126	-.84454
		4	-1.657108*	.097120	.000	-1.92340	-1.39082
		5	-2.347631*	.095596	.000	-2.60974	-2.08552
	3	1	1.963391*	.103348	.000	1.68003	2.24675
		2	1.127899*	.103348	.000	.84454	1.41126
		4	-.529209*	.067657	.000	-.71471	-.34370
		5	-1.219732*	.065450	.000	-1.39919	-1.04028
	4	1	2.492599*	.097120	.000	2.22631	2.75889
		2	1.657108*	.097120	.000	1.39082	1.92340
		3	.529209*	.067657	.000	.34370	.71471
		5	-.690523*	.055094	.000	-.84158	-.53946
	5	1	3.183122*	.095596	.000	2.92101	3.44523
		2	2.347631*	.095596	.000	2.08552	2.60974
		3	1.219732*	.065450	.000	1.04028	1.39919
		4	.690523*	.055094	.000	.53946	.84158

*. The mean difference is significant at the 0.05 level.

2. Descriptive statistics of the scores by skill domain of scripts from the joint study by level

Table AII-2-1: Descriptive statistics of the scores by skill domain of scripts from the joint study (Level 5)

Statistics

		1 Understanding	2 Information-handling	3 Perspectives	4 Synthesis	5 Evaluation	6 Cultures/values	7 Evidence	8 Communication	Overall Average
N	Valid	136	102	136	136	136	135	136	136	136
	Missing	225	259	225	225	225	226	225	225	225
Mean		4.45864	4.71078	4.42279	4.25110	3.66691	3.71676	4.25708	4.60993	4.24190
Std. Error of Mean		.045812	.042626	.044474	.055469	.060372	.058596	.053681	.036745	.037358
Std. Deviation		.534256	.430497	.518647	.646874	.704050	.680830	.626026	.428512	.435669
Variance		.285	.185	.269	.418	.496	.464	.392	.184	.190
Skewness		-.784	-1.247	-.397	-.204	-.084	-.102	-.242	-.784	-.392
Std. Error of Skewness		.208	.239	.208	.208	.208	.209	.208	.208	.208
Percentiles	25	4.00000	4.50000	4.00000	3.60000	3.52500	3.60000	3.60000	4.25000	3.94007
	50	4.50000	5.00000	4.25000	4.30000	3.60000	3.60000	4.30000	4.75000	4.27857
	75	5.00000	5.00000	5.00000	5.00000	3.97188	3.97500	5.00000	5.00000	4.64911

Table AII-2-2: Descriptive statistics of the scores by skill domain of scripts from the joint study (Level 4)

Statistics

		1 Understanding	2 Information-handling	3 Perspectives	4 Synthesis	5 Evaluation	6 Cultures/values	7 Evidence	8 Communication	Overall Average
N	Valid	112	84	112	112	111	109	112	112	112
	Missing	249	277	249	249	250	252	249	249	249
Mean		3.76116	4.19643	3.80804	3.48683	2.94414	2.99610	3.36763	3.97321	3.55138
Std. Error of Mean		.057374	.069197	.056810	.053131	.067845	.064163	.059219	.043519	.043948
Std. Deviation		.607191	.634205	.601221	.562291	.714787	.669881	.626720	.460558	.465100
Variance		.369	.402	.361	.316	.511	.449	.393	.212	.216
Skewness		-1.278	-.125	-1.509	-.368	-.253	-.367	-.436	-.896	-.589
Std. Error of Skewness		.228	.263	.228	.228	.229	.231	.228	.228	.228
Percentiles	25	3.50000	4.00000	3.50000	3.22500	2.30000	2.30000	3.20000	4.00000	3.32031
	50	4.00000	4.00000	4.00000	3.60000	2.97500	3.27500	3.60000	4.00000	3.59219
	75	4.00000	5.00000	4.00000	3.60000	3.60000	3.60000	3.60000	4.18750	3.86228

Table AII-2-3: Descriptive statistics of the scores by skill domain of scripts from the joint study (Level 3)

Statistics

		1 Understanding	2 Information-ing	3 Perspectives	4 Synthesis	5 Evaluation	6 Cultures/values	7 Evidence	8 Communication	Overall Average
N	Valid	40	30	40	40	39	36	40	40	40
	Missing	321	331	321	321	322	325	321	321	321
Mean		3.28750	3.73333	3.15625	3.04938	2.51731	2.80972	3.09688	3.50000	3.12177
Std. Error of Mean		.111283	.130171	.114175	.109187	.122334	.119619	.123167	.096410	.090916
Std. Deviation		.703813	.712975	.722104	.690561	.763974	.717717	.778974	.609750	.575005
Variance		.495	.508	.521	.477	.584	.515	.607	.372	.331
Skewness		-.967	-.759	-.861	.045	-.578	-.629	-.158	-.413	-.277
Std. Error of Skewness		.374	.427	.374	.374	.378	.393	.374	.374	.374
Percentiles	25	3.00000	3.18750	3.00000	2.30000	2.30000	2.30000	2.30000	3.00000	2.73214
	50	3.12500	4.00000	3.00000	3.26250	2.30000	2.75000	2.95000	3.50000	3.22500
	75	4.00000	4.00000	3.75000	3.60000	2.97500	3.55000	3.60000	4.00000	3.55625

3. Correlation statistics of the scores by skill domain (Levels 1 to 5)

Table AII-3: Correlation statistics of the scores by skill domain (Levels 1 to 5)

Correlations

		1 Understanding	2 Information-handling	3 Perspectives	4 Synthesis	5 Evaluation	6 Cultures/values	7 Evidence	8 Communication
1 Understanding	Pearson Correlation	1	.794**	.946**	.890**	.816**	.830**	.863**	.929**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
	N	360	252	360	360	358	345	360	359
2 Information-handling	Pearson Correlation	.794**	1	.789**	.763**	.628**	.619**	.775**	.826**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
	N	252	252	252	252	250	238	252	252
3 Perspectives	Pearson Correlation	.946**	.789**	1	.902**	.810**	.801**	.865**	.928**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
	N	360	252	360	360	358	345	360	359
4 Synthesis	Pearson Correlation	.890**	.763**	.902**	1	.814**	.768**	.870**	.888**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
	N	360	252	360	360	358	345	360	359
5 Evaluation	Pearson Correlation	.816**	.628**	.810**	.814**	1	.781**	.779**	.800**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000

	N	358	250	358	358	358	344	358	357
6	Pearson Correlation	.830**	.619**	.801**	.768**	.781**	1	.788**	.813**
Cultures/Values	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
	N	345	238	345	345	344	345	345	344
7 Evidence	Pearson Correlation	.863**	.775**	.865**	.870**	.779**	.788**	1	.873**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
	N	360	252	360	360	358	345	360	359
8	Pearson Correlation	.929**	.826**	.928**	.888**	.800**	.813**	.873**	1
Communication	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	
	N	359	252	359	359	357	344	359	359

**. Correlation is significant at the 0.01 level (2-tailed).

4. The complete output of ANOVA statistics of the average scores of Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values by level

4A: The complete output of ANOVA statistics of the overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values by candidate of the joint study (Levels 3 to 5)

Table AII-4A-1: Descriptives of the overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values by candidate of the joint study (Levels 3 to 5)

Descriptives

Overall Average

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
3	10	3.14224	.247829	.078370	2.96495	3.31953	2.679	3.483
4	28	3.56858	.201608	.038100	3.49041	3.64676	3.166	3.887
5	34	4.26112	.248638	.042641	4.17437	4.34788	3.620	4.661
Total	72	3.83640	.484265	.057071	3.72260	3.95020	2.679	4.661

Table AII-4A-2: ANOVA statistics of overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values by candidate of the joint study (Levels 3 to 5)

ANOVA

Overall Average

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	12.960	2	6.480	121.162	.000
Within Groups	3.690	69	.053		
Total	16.650	71			

Table AII-4A-3: Statistics of Post Hoc Tests of overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values by candidate of the joint study (Levels 3 to 5)

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Overall Average

Tukey HSD

(I) Level	(J) Level	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
3	4	-.426343*	.085196	.000	-.63041	-.22227
	5	-1.118882*	.083194	.000	-1.31816	-.91961
4	3	.426343*	.085196	.000	.22227	.63041
	5	-.692539*	.059018	.000	-.83390	-.55117
5	3	1.118882*	.083194	.000	.91961	1.31816
	4	.692539*	.059018	.000	.55117	.83390

*. The mean difference is significant at the 0.05 level.

4B: The complete output of ANOVA statistics of the overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values (Levels 1 to 5) by level (Levels 1 to 5)

Table AII-4B-1: Descriptives of the overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values (Levels 1 to 5)

Descriptives

Average of Domains 4_5_6

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	24	1.000	.000	.000	1.000	1.000	1.000	1.000
2	24	1.686	.336	.069	1.544	1.828	1.433	2.300
3	64	2.731	.567	.071	2.590	2.873	1.433	3.883
4	112	3.142	.553	.052	3.038	3.245	1.500	4.533
5	136	3.876	.544	.0467	3.784	3.969	2.442	5.000
Total	360	3.106	.975	.051	3.005	3.207535940000000	1.000000000000000	5.000000000000000

Table AII-4B-2: ANOVA statistics of the overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values (Levels 1 to 5)

ANOVA

Average of Domains 4_5_6

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	244.676	4	61.169	224.297	.000
Within Groups	96.813	355	.273		
Total	341.489	359			

Table AII-4B-3: Statistics of Post Hoc Tests of the overall average of scores for Skill Domains 4 Synthesis, 5 Evaluation and 6 Cultures/Values (Levels 1 to 5)

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Average of Domains 4_5_6

Tukey HSD

(I) Level	(J) Level	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.686	.151	.000	-1.100	-.273
	3	-1.731*	.125	.000	-2.074	-1.389
	4	-2.142*	.117	.000	-2.464	-1.820
	5	-2.876*	.116	.000	-3.193	-2.559
2	1	.686*	.151	.000	.273	1.099
	3	-1.045*	.125	.000	-1.388	-.702
	4	-1.456*	.117	.000	-1.778	-1.133
	5	-2.190*	.116	.000	-2.507	-1.873
3	1	1.731*	.125	.000	1.389	2.074
	2	1.045*	.125	.000	.702	1.388
	4	-.410*	.082	.000	-.635	-.186
	5	-1.145*	.079	.000	-1.362	-.928

4	1	2.142*	.117	.000	1.820	2.464
	2	1.456*	.117	.000	1.133	1.778
	3	.410*	.082	.000	.186	.635
	5	-.735*	.067	.000	-.917	-.552
5	1	2.876*	.116	.000	2.559	3.193
	2	2.190*	.116	.000	1.873	2.507
	3	1.145*	.0792	.000	.928	1.362
	4	.735*	.067	.000	.552	.917

*. The mean difference is significant at the 0.05 level.

5. Statistics of the by-question scores for Domains 5 Evaluation and 6 Cultures/Values

Table AII-5-1: Descriptives of the by-question scores for Domains 5 Evaluation and 6 Cultures/Values

Descriptives

						95% Confidence Interval for Mean			
		N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
5 Evaluation	1	83	2.82997	1.133272	.124393	2.58251	3.07743	1.000	5.000
	2	83	3.18705	.997599	.109501	2.96922	3.40488	1.000	5.000
	3	84	2.82545	.941057	.102678	2.62122	3.02967	1.000	5.000
	4	108	2.84282	1.070745	.103033	2.63857	3.04707	1.000	4.650
	Total	358	2.91557	1.046640	.055317	2.80679	3.02436	1.000	5.000
6 Cultures/values	1	76	2.91530	.967423	.110971	2.69423	3.13636	1.000	5.000
	2	81	3.21512	1.099188	.122132	2.97207	3.45817	1.000	5.000
	3	81	2.90895	1.009945	.112216	2.68563	3.13227	1.000	5.000
	4	107	2.87921	1.108248	.107138	2.66679	3.09162	1.000	4.575
	Total	345	2.97301	1.057799	.056950	2.86099	3.08502	1.000	5.000

Table AII-5-2: The ANOVA statistics for the by-question scores for Domains 5 Evaluation and 6 Cultures/Values

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
5 Evaluation	Between Groups	7.979	3	2.660	2.458	.063
	Within Groups	383.099	354	1.082		
	Total	391.078	357			
6 Cultures/values	Between Groups	6.275	3	2.092	1.884	.132
	Within Groups	378.640	341	1.110		
	Total	384.915	344			

Table AII-5-3: The Post Hoc Tests for the by-question scores for Domains 5 Evaluation and 6 Cultures/Values

Post Hoc Tests

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) Question	(J) Question	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
5 Evaluation	1	2	-.357078	.161484	.122	-.77391	.05976
		3	.004523	.161003	1.000	-.41107	.42012
		4	-.012854	.151852	1.000	-.40483	.37912
	2	1	.357078	.161484	.122	-.05976	.77391
		3	.361602	.161003	.113	-.05399	.77719
		4	.344224	.151852	.108	-.04775	.73620
	3	1	-.004523	.161003	1.000	-.42012	.41107
		2	-.361602	.161003	.113	-.77719	.05399
		4	-.017378	.151340	.999	-.40803	.37327
	4	1	.012854	.151852	1.000	-.37912	.40483
		2	-.344224	.151852	.108	-.73620	.04775
		3	.017378	.151340	.999	-.37327	.40803

Table AIII-1a Excerpts of different types of *Knowledge* in Sample A (Level 5)

Level 5 (Sample A from Candidate A)		
Type of Knowledge	**Perspective**	**Excerpt**
Fact	Local (Economic)	*L5.1: 'the economy in Hong Kong (was) boosted … under the Individual Visit Scheme' (P1Q3a)*
	Local (Socio-political)	*L5.2: 'For instance, the previous Chief Editor of Ming Pao (a local newspaper), Mr Lau Chun-to had been attacked and injured seriously.' (P2Q1a)* *L5.3: 'For example, the news of TVB has been criticized as self-censored.' (P2Q1a)* *L5.4: 'For example, (the) Chief Executive CY Leung has issued a legal letter to the Chief Editor of a Hong Kong local newspaper, for the newspaper's editorial criticizing Mr Leung' (P2Q1a)*
	Global (Cultural)	*L5.5: 'For instance, people in Thailand believe that touching people's head is impolite' (P1Q3b)*
Generalisation	Personal	*L5.6: 'adolescents at this age are still at a stage of seeking (for) peer recognition' (P1Q2a)*
	Local (Socio-political)	*L5.7: 'Under a stable and harmonious society, press will report news freely without fear' (P2Q1a)* *L5.8: 'With (a) high degree of press freedom, media will discuss governance problem or policies freely on TV or newspaper, encouraging socio-political participation.' (P2Q1b)*
	National (Cultural)	*L5.9: 'Chinese people depend highly on agricultural products like rice in daily dining' (P1Q1b)*
	National (Social)	*L5.10: '…it is impossible for the rural population to have a high living standard as that of urban population.' (P1Q1a)* *L5.11: 'Under the current household registration system, they have less social welfare…' (P1Q1b)*
	National (Social-political)	*L5.12: 'By legislation, such behavior can be monitor(ed) and prohibited' (P1Q1c)*
	Global (cultural)	*L5.13: 'under the flow of western culture of individualism and freedom, plastic surgery has become very common and acceptable in the society' (P1Q2a)* *L5.14: 'It is extremely common for international tourists to travel through airplanes.' (P1Q3b)*
Principle	Personal	*L5.15: '…with the promotion of celebrities through mass media, the social norm change(s) to accept plastic surgery and believe that it a way to boost self-esteem' (P1Q2a)*
	Local (Socio-political)	*L5.16: 'a comprehensive legal system and legal independence ensure the press freedom and prevent government intervention.' (P2Q1a)* *L5.17: 'press freedom helps expressing (express) social discontent, improving social harmony' (P2Q1b)*
	National (Social)	*L5.18: '…urbanization of China, leading to the decreasing number of farmers and (in) rural area(s).' (P1Q1a)* *L5.19: 'social disharmony arised (caused) by urban-rural disparity and migrant workers' (P1Q1b)*
	National (Socio-political)	*L5.20: 'the government can provide subsidy in terms of agricultural product quantity as to raise their living standard' (P1Q1c)*
	National (Environmental)	*L5.21: '…over-reliance on chemical fertilisers or pesticides and heavy metal pollution, the quality of arable land drop(s)' (P1Q1b)*
	Global (Economic)	*L5.22: 'related industries will be boosted such as transport or factories producing souvenirs.' (P1Q3a)*
	Global (Environmental)	*L5.23: 'Since negative consequences led by global warming like extreme weather or rising sea level…' (P1Q3b)*
	Global (Social)	*L5.24: 'These conflicts will lead to growing dissatisfaction towards tourists or even damaged the cultural or historical relics.' (P1Q3b)*

Table AIII-1b Excerpts of different types of *Knowledge* in Sample B (Level 4)

Level 4 (Sample B from Candidate B)		
Type of Knowledge	**Perspective**	**Excerpt**
Vocabulary term	Local (Socio-political)	*L4.1: 'ICAC' (P2Q1a)*
	National (Social)	*L4.2: '…a huge disparity between urban and rural (in terms of economy)' (P1Q1a)*
Fact	Local (Socio-political)	*L4.3: '…in Hong Kong three powers are separated…' (P2Q1a)* *L4.4: 'For example, a newspaper spread some negative opinions about CY Leung's (the Chief Executive) family and his ability.' (P2Q1a)* *L4.5: 'For instance, two transsexual feel unfair because they can't have marriage (get married). ... By voicing out their opinions by the mass media…' (P2Q1b)*
Generalisation	Personal	*L4.6: 'As young people in (at) this age wants to be independent, seek recognition and identity of others among the peers…' (P1Q2b)*
	Local (Socio-political)	*L4.8: 'Therefore, the mass media spread something negative to (about) the government, the government cannot sue the newspaper…' (P2Q1a)* *L4.9: 'When the government is doing anything or proposing measures, the mass media will keep checking it…' (P2Q1b)*
	National (Socio-economic)	*L4.10: '…the farmer will choose to leave the family to earn a living in urban (areas)' (P1Q1b)*
	Global (Economic)	*L4.11: '…international tourists, they need to aboard to airplane(s)' (P1Q3b)*
Principle	Personal	*L4.12: 'they will be teased or treated unequally in the society. Therefore, the(ir) self-esteem is low…' (P1Q2a)*
	Local (Socio-political)	*L4.13: 'First, the low level of corruption will improve the press freedom in Hong Kong.' (P2Q1a)* *L4.14: '…no official has the right to suppress it (press freedom) because the other powers will monitor this act.' (P2Q1a)*
	National (Environmental)	*L4.15: '…restricting the high amount of chemical sewage flowing to (the) river, farmers can get clear water and high quality land free from heavy metal(s)' (P1Q1c)*
	Global (Economic)	*L4.16: 'Some countries can develop tourism in their countries and develop different industries so the tourists can consume.' (P1Q3a)*
	Global (Environmental)	*L4.17: 'Airplane(s) will burn the fuels to release more CO_2, which results in global warming due to greenhouse effect.' (P1Q3b)*
Error	National (Social)	*L4.18: 'This will cause the problem of intergenerational family' (should be skipped generation family), as…the farmer will choose to leave the family to earn a living in urban. The kids and their parents will stay…' (P1Q1b)*

Table AIII-1c Excerpts of different types of *Knowledge* in Sample C (Level 3)

Level 3 (Sample C from Candidate C)		
Type of Knowledge	**Perspective**	**Excerpt**
Vocabulary term	Personal	*L3.1: 'physically mature' (P1Q2b)* *L3.2: 'mentally mature' (P1Q2b)* *L3.3: 'permissive parents' (P1Q2b)*
	Global (Social)	*L3.4: 'anti-globalisation' (P1Q3b)* *L3.5: 'global harmony' (P1Q3b)*
Fact	Local/ National (Socio-economic)	*L3.6: 'Wong Zhang (Huang Jing) of (the) Asian (Asia) Television …force(d) (the) Asian (Asia) Television to selectively broadcast news unfavourable to protesters' (P2Q1a)*
Generalisation	Personal	*L3.7: '…those who are not physically attractive may feel less self-worth(y) and can seek (gain) less recognition from (the) peers.' (P1Q2a)* *L3.8: '…they think plastic surgery is acceptable as celebrities accept it.' (P1Q2a)* *L3.9: '…they may make decision (decisions) too easily and is not mentally strong enough to deal with possible side effects and failure(s).' (P1Q2b)*
	Local (Socio-political)	*L3.10: '…if the boss has certain political stance, and he (may) interupt (interrupt) the operation of his own company…' (P2Q1a)* *L3.11: '…some mainland and real estate company (companies) withdraw advertisement or sponsorship to "punish" press for publishing news unfavourable to them.' (P2Q1a)* *L3.12: '…free press …make citizens have a better understanding about the policy' (P2Q1b)* *L3.13: '…press freedom let people know the dark side of (the) government, cultivate anger and discontent toward (the) government…drag back governance efficiency' (P2Q1b)*
	National (Socio-political)	*L3.14: '…implies farmers move to urban area(s)as migrant workers…' (P1Q1a)* *L3.15: '…grass root people …discontent with the government and the rich, dissatisfied about social inequality' (P1Q1b)*
	Global (Economic)	*L3.16: '…tertiary industries (industries) such as retailing, nations, earn more foreign exchange…' (P1Q3a)*
Principle	National (Environmental)	*L3.17: '…excessive use of chemical fertilisers and over consumed land cause contamination' (P1Q1b)* *L3.18: 'Capital needed to environmentally friendly products, this provide(s) finantial (financial) incentives for farmers to improve problems in farmland.' (P1Q1c)*
	National (Socio-economic)	*L3.19: 'imbalanced distribution of wealth cause(s) (a) wealth gap in rural and urban area further increased. Farmer(s) leave their homeland to urban cities…' (P1Q1b)*
Error	Local (Socio-political)	*L3.20: 'The Chief Executive has made a public statement to scold a student newspaper of (the) Chinese University of Hong Kong ' (Should be the University of Hong Kong) (P2Q1a)* *L3.21: 'free press act(s) as coordinators between (the) government and citizens' (P2Q1b)*
	National (Environmental)	*L3.22: 'Environmental problems also harm the mental health of residents living nearby' (P1Q1b)*
	National (Socio-economic)	*L3.23: 'immigrant workers' (should be 'migrant workers' (P1Q1b)* *L3.24: 'China's economy is shifting to secondary and tertiary industry, loss in foreign investment in primary industry is a bearable cost.' (P1Q1c)*
	Cultural	*L3.25: 'in a harmony perspective' (P1Q3b)*

Table AIII-1d Excerpts of different types of *Knowledge* in Sample D (Level 2)

Level 2 (Sample D from Candidate D)		
Type of Knowledge	**Perspective**	**Excerpt**
Vocabulary term	Personal	*L2.1: 'self-esteem' (P1Q2a)* *L2.2: 'conflicts' (P1Q3b)*
Fact	Environmental	*L2.3'over-exploited the underground water' (P1Q1b)*
	Local (Socio-political)	*L2.4: 'The government interferes' (P2Q1a)* *L2.5: '…the policy (will be) carried out …smoothly and effective(ly)…Thus, (the) effectiveness of governance is enhanced' (P2Q1b)*
Generalisation	Personal	*L2.6: 'Due to social mass media the youngsters are easily influence(d), without knowing truth behind these promotions…' (P1Q2a)* *L2.7: '…youngsters…are young and not mature enough to make decision(s)' (P1Q2b)* *L2.8: '…youngsters are quite fast at making their decisions and changing them too.' (P1Q2b)*
	Local (Socio-political)	*L2.9: '…the high degree of press freedom ensure(s) public speak out their opinions towards policies' (P2Q1b)*
Principle	National (Economic)	*L2.10: '…since the insufficiency of labour force, the number of food production would decrease seriously' (P1Q1b)*
Error	Local (Socio-political)	*L2.11: '…people start to be threaten(ed) due to the strict rule of law, the activities that damage the environment can be reduced…' (P1Q1c)* *L2.12: 'People have the right to speak up and it (is called) calls press freedom.' (P2Q1a)* *L2.13: '…platform on communication with government also affect press freedom in Hong Kong.' (P2Q1a)* *L2.14: 'A high degree of press freedom consolidate the confidence of public' (P2Q1b)*

Table AIII-1e Excerpts of different types of *Knowledge* in Sample E (Level 1)

Level 1 (Sample E from Candidate E)		
Type of Knowledge	**Perspective**	**Excerpts**
Vocabulary term	Personal	*L1.1: 'boost their confidence in social gatherings' (P1Q2a)*
	Social	*L1.2: 'discrimination' (P1Q1b)* *L1.3: 'health risks' (P1Q2b)*
	Local (Socio-political)	*L1.4: 'legislation…will be strictly forbiddening (forbidding) the teenagers' (P1Q2b)* *L1.5: 'The different media show different views… It is one of the way of citizens use their freedom of expression.' (P2Q1a)* *L1.6: 'rule of law' (P2Q1a)*
	Global (Environmental)	*L1.7: 'global warming' (P1Q3b)*
Fact	Local (Political)	*L1.8: 'In 2003, the Hong Kong government want(ed) to set up a law that call (is called) number 23 (Article 23)' (P2Q1a)*
Generalisation	Local (Socio-political)	*L1.9: '…sometime the government may limit or hide some news that is not good for the government' (P2Q1a)*
Error	Personal	*L1.10: '…they want to…stabilizing their appearance not only in their career paths, but also in their quality of life.' (P1Q2a)* *L1.11: '…plastic surgery…will be creating a lot of deaths on the teenagers.' (P1Q2b)*
	Social	*L1.12: 'living problems Rents in urban areas have been increasing' (P1Q1b)*
	Global (Political)	*L1.13: Tourism enhancing 'international image' (P1Q3a)*

Appendix III

Table AIII-2a Excerpts of Domain 6 *Cultures/Values* from Sample A (Level 5)

Level 5 (Sample A from Candidate A)		
		Excerpt
Culture	Difference in the way of life	*L5.1: 'social disharmony arised by (caused by) urban-rural disparity and migrant workers' (P1Q1b)*
	Conflict	*L5.2: the second concern is the cultural conflicts arised from (caused by) international tourism.' (P1Q3b)*
Value	Personal values	*L5.3: 'Secondly, in terms of addressing the root problem, passing law is not dealing with the root cause of teenagers undergoing plastic surgery which is gaining peer recognition and incorrect values toward beauty.' (P1Q2b)*
	Economic values	*L5.4: 'If the government intervene(s) (with) the free market by passing laws to bar business opportunities, (the) profit of these companies may drop and they may oppose to the government' (P1Q2b)*
	Social norm	*L5.5: 'However, with the promotion of celebrities through mass media, the social norm change(s) to accept plastic surgery and believe that it is a way to boost self-esteem…' (P1Q2a)*
	Press freedom	*L5.6: 'Press freedom include(s) freedom of expressing ideas or reporting news, be it positive or negative' (P2Q1a)*
	Social harmony	*L5.7: '…press freedom help(s) expressing (express) social discontent, improving social harmony' (P2Q1b)*
	Civil values, civil awareness	*L5.8: '(a) high degree of press freedom can play the role of educator or promotor, helping the government to inculcate correct civil values on citizens, raising civil awareness.' (P2Q1b)*
Value Position	Counter-argument	*L5.9: 'Someone may also argued that high degree of press freedom is actually hindering government because it leads to and encourages demonstrations or strikes, and are opposing the government, which will then worsen relationship between the government and Hong Kong people….* *However, such negative or dark side of government being reported is actually helping to improve the policy and government performance. For example, when the mass media reveal some negative side of a policy, the government can then fix the problem. Actually, the quality of policy is more important than it is not criticised by the public in terms of governance' (P2Q1b)*

Table AIII-2b Excerpts of Domain 6 *Cultures/Values* from Sample B (Level 4)

Level 4 (Sample B from Candidate B)		
		Excerpt
Culture	Difference in the way of life	*L4.1: 'there will be disparity between urban area and rural area' (P1Q1b)*
	Cultural difference	*L4.2: 'While different countries have different culture(s), there will be argument(s) and conflicts between locals and tourists.' (P1Q3b)* *L4.3: 'Sometimes visitors in India use left hand to touch an Indian, which is disrespectful in Indian culture, but generally not in other cultures.' (P1Q3b)*
Value	Personal values	*L4.4: 'As some people are too ugly, they will be teased or treated unequally in the society.' (P1Q2a)* *L4.5: 'Plastic surgery is now generally well accepted surgery in society as it doesn't do harm to others, while having few benefits to the teenager(s) himself.' (P1Q2b)*
	Rights	*L4.6: 'the right of the press is well protected by law.' (P2Q1a)*
Value Position	Counter-argument	*L4.7: 'Some says plastic surgery have risk and will kill the people or will fail.' (P1Q2b)* *L4.8: 'Some say that a high press freedom may hinder the implementation of policies as the negative opinion spreaded (spread) will always draw attention of certain stakeholders to resist he decision of government. However, this procedure is in fact help(s) the government to understand the opinions of different stakeholders. So the government can adjust their measures or explain publicly with reasons so as to meet the demands of different stakeholders' (P2Q1b)*

Table AIII-2c Excerpts of Domain 6 *Cultures/Values* from Sample C (Level 3)

Level 3 (Sample C from Candidate C)		
		Excerpt
Culture	Difference in the way of life	*L3.1: 'imbalanced distribution of wealth causes wealth gap in rural and urban area further increased' (P1Q1b)*
Value	Personal values	*L3.2: '…showing this generation use outer appearance to judge people' (P1Q2a)* *L3.3: 'Under-18s are not mentally mature to take the surgery as they are still developing their values and critical thinking ability, as the surgery is permanent…as they may make decition (decisions) too easily and is not strong enough to deal with possible side effects and failure...' (P1Q2b)*
Value Position	Counter-argument	*L3.4: 'Some may argue this force(s) under-18s to go to underground operating rooms, and increase the risk of operation.' (P1Q2b)* *L3.5: 'People may argue that press freedom let people know the dark side of (the) government, cultivate anger and discontent toward (the) government, hence make (making) social movements happens (happen) more frequently and drag(ging) back governance efficiency.' (P2Q1b)*

Table AIII-2d Excerpts of Domain 6 *Cultures/Values* from Sample D (Level 2)

Level 2 (Sample D from Candidate D)		
		Excerpt
Culture	Respect, conflict	*L2.1: '…tourists do not respect the traditions… and may cause conflicts' (P1Q3b)*
Value	Personal values	*L2.2: 'However, they tend to mature and would like to make their own decisions, since they are still at that age, where they are identifying themselves and they are easily influence(d) by the mass media.' (P1Q2b)*

Table AIII-2e Excerpts of Domain 6 *Cultures/Values* from Sample E (Level 1)

Level 1 (Sample E from Candidate E)		
		Excerpt
Culture	Difference in way of life	*L1.1: '…residents who left rural are usually lack of skills, not to say expertise.' (P1Q1b)*
	Respect	*L1.2: 'for the concern of respecting one another culture' (P1Q3b)*
Value	Personal values	*L1.3: '…undergo the plastic surgery it is because that they want to boost their confidence in social gatherings and to facilitate their personal images in order to get in close with their friends and to rebuild their relationships' (P1Q2a)*
Value Position	Counter-argument	*L1.4: 'Some of the people may say were view will make a lot of argue(ments) when there is a high degree of press freedom. But I think people will look for the compromise. The compromise can help the government to solve problem(s). And it is a good choice to increase the credibility of the government' (P2Q1b)*

Appendix III

Table AIII-3a Excerpts of Domain 2 *Information-handing* from Sample A (Level 5)

Level 5 (Sample A from Candidate A)	
	Excerpt
Generalise trend	*L5.1: 'From Source A, contribution of primary industry, including farming, drop(ped) from 27.1% in 1990 to 10% in 2003, drop(ped) by 17.1% in 23 years.' (P1Q1a)* *L5.2: '…percentage of rural population drop(ped) from 73.6% in 1990 to 46.3% in 2013 as shown in Source B, showing that less Chinese citizens work in farms or in the agricultural industry.' (P1Q1a)* *L5.3: 'Economic contribution, which is one tourism receipts also rise (rose) from 262 billion US dollars in 1990 to 1078 billion in 2012, the number has increased by almost 5 times while that of tourist number has rise(n) by more than 2 times.' (P1Q1a)* *L5.4: 'In Source A, (the) number of international tourist s arrived rise (rose) from 4.34 million in 1990 to 1035 million in 2012, showing a continuous rising trend in the 2 decades. Economic contribution… rised (rose)… the number has increased by almost 5 times while that of tourist number has rise(n) more than 2 times.' (P1Q3a)*
Interpret information	*L5.5: 'In Source C, urban people are in an express railway while rural people are chasing after in a horse, which implies that it is impossible for the rural population to have a high living standard on that of urban population.' (P1Q1a)*

Table AIII-3b Excerpts of Domain 2 *Information-handing* from Sample B (Level 4)

	Level 4 (Sample B from Candidate B)
	Excerpt
Generalise trend	*L4.1: 'From Source A, we can see that the percentage contribution to GDP in China by primary industry …is decreasing gradually from 27.1% in 1990 to 10.0% in 2013.' (P1Q1a)* *L4.2: 'In the source, both tourist arrivals and tourism receipts increase sharply from 1990 to 2012. The arrivals increase from 434 million in 1990 to 1035 million in 2012, while the receipts increase from 262 million to 1078 million in (the) same period of time.' (P1Q3a)* *L4.3: 'From Source B, we can find that the CO_2 emission due to air transport is nearly triple in 2035, while it was 43% in 2005.' (P1Q3b)*
Interpret information	*L4.4: 'From Source C, we can see that the income of urban residents is like a train, while that of farmer is like a wagon.' (P1Q1a)*

Table AIII-3c Excerpts of Domain 2 *Information-handing* from Sample C (Level 3)

	Level 3 (Sample C from Candidate C)
	Excerpt
Generalise trend	*L3.1: '… the percentage GDP decrease(s) drastically from 27.1% at 1990 to 10% in 2013, while that of secondary industry increase(s) slightly and that of tertiary industry increase(s) largely.' (P1Q1a)* *L3.2: '…from Source B, the percentage of rural population decreased drasticly (drastically) from 73.6% in 1990 to 46.3% in 2013, the percentage almost decreased by half…' (P1Q1a)* *L3.3: 'the international tourist arrivals increased from 434 million in 1990 to 1035 million at 2013, the increase reach(ed) 150%, having a drastic increase in tourist while international tourism receipts increased from 262 billion USD to 1078 billion USD, the increase reach(ed) about 300%, showing tourism industry is decreasing rapidly.' (P1Q3a)* *L3.4: '…from Source B, emission of CO_2 caused by tourism increase(d) by 200% over 200 years' (P1Q3b)*
Interpret information	*L3.5: '…from Source B, the cartoon shows the income of urban residents is progressing like a train while that of farmers move(s) forward like a slow horse cart.' (P1Q1a)*

Table AIII-3d Excerpts of Domain 2 *Information-handing* from Sample D (Level 2)

	Level 2 (Sample D from Candidate D)
	Excerpt
Generalise trend	*L2.1: 'According to Source A, the percentage contribution of primary industry were (was) decreasing gradually from 27.1% to 10.0% between 1990 and 2013, the percentage contribution of tertiary industry were (was) increasing gradually from 31.5% to 46.1%.' (P1Q1a)* *L2.2: 'According to Source B, the percentage of rural population in the total population of China were decreased over the half, from 73.6% to 46.3% between 1990 and 2013.' (P1Q1a)* *L2.3: 'The international tourist arrivals shown in Source A is increasing sharply from 434 million people in 1990 to 1035 million people in 2012.' (P1Q3a)* *L2.4: 'In Source B, it shows that the carbon dioxide emission from air transport is the largest in both 2005 (43%) and 2035 (53%). And it shows the carbon dioxide emission from air transport is increasing sharply by 43% in 2005 to 53% in 2035.' (P1Q3b)*
Interpret information	*L2.5: 'In Source B, the overall carbon dioxide emission in 2035 is about 3000 million tonnes of carbon dioxide emission that is much more than that in 2005, which is no more than 1500 million tonnes.' (P1Q3b)*

Table AIII-3e Excerpts of Domain 2 *Information-handing* from Sample E (Level 1)

	Level 1 (Sample E from Candidate E)
	Excerpt
Generalise trend	*L1.1: 'With reference to Source A, there has been an overall decrease from 27.1% in 1990 to 100% in 2013, nearly by two thirds.' (P1Q1a)*
Interpret information	*L1.2: 'With reference to Source C, from the picture which the two men travelling by train and caravan labelled "income" respectively, indicate (indicating) that wealth disparity has been further widened.' (P1Q1a)* *L1.3: 'For the international tourist arrivals, it has increased from 434 million in 1990 to 1035 million in 2012, while for the international tourism receipts, it increased from 262 billion dollars in 1990 to 1075 billion dollars in 2012.' (P1Q3a)* *L1.4: 'It shows that, if the international tourism keep(s) on increasing, the comparison of CO_2 will follow to increase.' (P1Q3b)*

Appendix III

Table AIII-4a Excerpts of *Formulation of Viewpoints, Opinions and Suggestions* from Sample A (Level 5)

Level 5 (Sample A from Candidate A)	
	Excerpt (from Paper 1 Q1(c))
Evaluation	*P1Q1L5.1:'So providing subsidy can solve the problem of social disharmony as migrant workers drop and living standard of farmers rised(rose).'*
	Excerpt (from Paper 1 Q2(b))
Synthesis	*P1Q2L5.1: 'Passing laws can only lead to behavioural change but not attitude change. In (the) long run, when the youngster(s) reach 18 years old, they will still undergo plastic surgery.'*
Evaluation	*P1Q2L5.2: 'in terms of addressing the root problem, passing law is not dealing with the root cause of teenagers undergoing plastic surgery which is gaining peer recognition and incorrect values toward beauty.'*
Counter-argument	*P1Q2L5.3: 'Someone may argue that the policy can take effect in short time to stop teenagers from undergoing plastic surgery, so as to prevent any medical accidents or negative consequence on the growth of teenagers in Hong Kong.'*
Rebuttal	*P1Q2L5.4: 'However, in (the) long run, it is not effective to change their values towards beauty and raise their awareness towards danger of such invasive procedure in surgeries. In fact, to solve such problem involving value judgement, soft measures should be used for long term effectiveness.'*
	Excerpt (from Paper 1 Q3(b))
Synthesis	*P1Q3L5.1: 'In fact, these behavior(s) may be accepted by mainlanders themselves, but not to Egyptians and Hong Kongers. Thus, conflicts arised (arose) due to cultural difference.'*
Evaluation	*P1Q3L5.2: 'Since negative consequences led by global warming like extreme weather or rising sea level is threatening the whole world, different countries will have the concern on carbon emission resulted in(from) internal tourism.'*
	P1Q3L5.3: 'So under these cultural conflicts, it will be a great concern for governments all over the world to try to reduce locals' dissatisfaction, educating tourists in adapting to local culture without sacrificing the economic benefit generated by international tourism.'
	Excerpt (from Paper 2 Q1(b))
Evaluation (reviewing government policies)	*P2Q1L5.1: 'Secondly, press freedom can help reviewing government policies, improving the quality of government policies, improving the quality of government policies. For example, media will invite specialist(s) to express their view toward important policies like Third Runway or Solid Waste Charging in Hong Kong, including both positive and negative comments.'*
	P2Q1L5.2: 'Under high degree of press freedom, media will voice out negative side of policies or even suggest for improvement. Thus, the quality of policy will be improved.'
	P2Q1L5.3: 'Since the effectiveness of governance does not simply lies on the feasibility of government policies, but also whether these policies can reach or satisfy citizens' demand, and so high degree of press freedom also (with) social reflection on important issues, help the government to formulate better policies and to reach consensus in the society easier.'
Counter-argument	*P2Q1L5.4: 'Someone may argue that high degree of press freedom is actually hindering governance because it leads to and encouraged demonstration or strikes, and are opposing the government, which will then worsen relationships between the government and Hong Kong People.'*
	P2Q1L5.5: 'For example, each year, Apple Daily will encourage people to join the strike on 1st July.'
Rebuttal	*P2Q1L5.6: 'However, such negative or dark side of government being reported is actually helping to improve the policy and government performance. For example, when the mass media reveal some negative side of a policy, the government can then fix the problem'*
	P2Q1L5.7: 'Actually, the quality of policy is more important than it is not criticized by the public in terms of governance.'
	P2Q1L5.8: 'For example, the policy of building 85000 housing flat(s) suggested by former Chief Executive Tung Chee-hwa, though is enforced and implemented, face serious criticize (criticism) and opposition after that, leading to the resignation of him. So the quality of policy is more important.'

Appendix III

Table AIII-4b Excerpts of *Formulation of Viewpoints, Opinions and Suggestions* from Sample B (Level 4)

Level 4 (Sample B from Candidate B)	
	Excerpt (from Paper 1 Q1(c))
Synthesise	*P1Q1L4.1:'with more advanced equipments, each farmer can produce more products and food with limited time, so they can sell their extra products after household responsibility system and earn more money for a living and buy food.'*
	P1Q1L4.2: 'Therefore, by strengthening the law to control the chemical pollution to rural areas by, for instance, restricting the high amount of chemical sewage flowing to river(s)…'
	Excerpt (from Paper 1 Q2(b))
Synthesise	*P1Q2L4.1: 'It is a freedom of teenagers to perform the plastic surgery.'*
Counter-argument	*P1Q2L4.2: 'Some says plastic surgery have risk and will kill the people or will fail.'*
Rebuttal	*P1Q2L4.2: 'It is the problem of the company and surgeon but not the surgery itself. The government can enhance the supervision of plastic surgery so that the surgery can succeed with a very high percentage*
	Excerpt (from Paper 1 Q3(b))
Synthesise	*P1Q3L4.1: 'Therefore, the increase of amount of tourists may intensify global warming and cause destruction to (the) environment.'*
	P1Q3L4.2: 'As the amount of tourists increase, more people will visit other countries. While different countries have different culture, there will be argument and conflicts between locals and tourists… So this cause(s) destruction to the heritage, arousing discontent of locals as the heritage are (is) their national properties.'
	Excerpt (from Paper 2 Q1(b))
Synthesise	*P2Q1L4.1: 'First of all, high degree of press freedom help(s) supervise the government. When the government is doing anything or proposing measures, the mass media will keep checking of (on) it and spread it to the public. Therefore, when the government try (tries) to do something which hinders the interest of certain stakeholder, the mass media will disclose this and the public may respond to it.'*
	P2Q1L4.2: 'For example, Wong Wai Kei is not allowed to get the permission to run his television programmes (broadcasting company) publicly in televisions due to a series of factors, in which the government didn't disclose because it's confidential. The mass media record it and spread it to the public. Therefore, the public think that the acts of government are not transparent enough.'
	P2Q1L4.3: 'This arouse(s) discontent of the public and people oppose the government, forcing her to give a detail explanation.'
Counter-argument	*P2Q1L4.4: 'Some say that a high press freedom may hinder the implementation of policies as the negative opinion spreaded (spread) will always draw attention of certain stakeholders to resist the decision of government.'*
Rebuttal	*P2Q1L4.5: 'However, this procedure is in fact help(s) the government to understand the opinions of different stakeholders. So the government can adjust their measures or explain publicly with reasons so as to meet the demands of different stakeholders in public.'*
	P2Q1L4.6: 'So the policy can have a higher conformity and implement with less resistance.'
	P2Q1L4.7: 'While people generally have a higher education level, they have critical thinking able to judge rationally.'

Appendix III

Table AIII-4c Excerpts of *Formulation of Viewpoints, Opinions and Suggestions* from Sample C (Level 3)

Level 3 (Sample C from Candidate C)	
	Excerpt (from Paper 1 Q1(c))
Synthesise	*P1Q1L3.1:'set higher import tax targeting imported food. So as to raise the selling price of imported food and increase the competitiveness of local farm products.'*
Counter-argument	*P1Q1L3.2:' '…people may say this hit the foreign trade market and lower the incentives of foreign investments.'*
Rebuttal	*P1Q1L3.3:' 'However, China economy is shifting to secondary and tertiary industry, loss in foreign investment in primary industry is a bearable cost.'*
	Excerpt (from Paper 1 Q2(b))
Synthesise	*P1Q2L3.1:' 'First, under-18s are not physically mature to take the surgery. Their body(ies) are still developing, there is a high chance of position shifting of inputted materials…'*
	P1Q2L3.2:' '…they may not be able to bear the negative impacts of surgery, as they may make decision(decisions) too easily and is not mentally strong enough to deal with possible side effects and failure.'
Counter-argument	*P1Q2L3.3:' 'Some may argue this force(s) under-18s to go to underground operating rooms and increase(s) the risk of operation.'*
Rebuttal	*P1Q2L3.4:' 'However, if (the) government do(es) its job as monitor, this would not happen.'*
	Excerpt (from Paper 1 Q3(b))
Synthesise	*P1Q3L3.1:' 'This intensify(ies) the problem of global warming and may result in more frequence(y) of natural hazards.'*
	P1Q3L3.2:' 'This intensify(ies) conflicts between nations or regions, may even cause anti-globalization movement or movement anti-tourists from certain country. This is bad to global harmony and lowering national hatred to each other.'
	Excerpt (from Paper 2 Q1(b))
Synthesise (Compare)	*P2Q1L3.1: 'Second, in terms of smoothly implemented, free press act as coordinators between government and citizens , explain the policy to citizens to make citizens have a better understanding about the policy, and can analysis benefit and cost themselves rationally.'*
	P2Q1L3.2: 'With low press freedom, citizens lost trust about "facts" on press and media lost function as coordinator and cannot help smooth implementation of policies.'
Counter-argument	*P2Q1L3.3: 'People may argue that press freedom let people know the dark side of government, cultivate anger and discontent toward government, hence make social movements happens (happen) more frequently and drag back governance efficiency.'*
Rebuttal	*P2Q1L3.4: 'This is true but it should be noticed that this happens only at the early stage desition (decision making) of governance.'*
	P2Q1L3.5: 'The effect is short. Once (the) government amend(s) (the) policies, social movement would stop.'
	P2Q1L3.6: 'But with low freedom of press, the negative effects on governance efficiency can be long lasting, such as making ineffective policies, having government official with low working abilities.'

Table AIII-4d Excerpts of *Formulation of Viewpoints, Opinions and Suggestions* Sample D (Level 2)

Level 2 (Sample D from Candidate D)	
	Excerpt (from Paper 1 Q1(c))
Synthesise	*P1.Q1L2.1:'…the government can enhance the inspection of human activities…. When (the) government use(s) the method of legislation to limit the activities, people start to be threaten due to the strict rule of law, the activities that damage the environment can be reduced …'*
	P1.Q1L2.2: 'If the government can promote the significance of conservation, people start to think about what they have done, change their attitude and not to do this activities anymore.'
	Excerpt (from Paper 1 Q2(b))
Synthesise	*P1.Q2L2.1: '…since they are still at that age, where they are identifying themselves and they can be easily influence(d) by the mass media.'*
	P1.Q2L2.2: 'If the government ban(s) these 'medical treatments' for youngsters, this will give them some time to rethink over their decision for plastic surgery, since at that age, youngsters are quite fast at making their decisions and changing them too.'
	Excerpt (from Paper 1 Q3(b))
Synthesise	*P1.Q3L2.1: 'Another global concern is that tourists do not respect the traditions and the rules made by the countries they travel and may cause conflicts between local people and tourists.'*
	Excerpt (from Paper 2 Q1(b))
Synthesise	*P2Q1L2.1: 'a high degree of press freedom lead(s) to less amendment after the policy has (is) carried out.'*
	P2Q1L2.2: 'Although, HKSAR allows public discussion during the consultation stage but most public cannot ensure their voice are actually being heard.'
	P2Q1L2.3: 'Thus, the high degree of press freedom ensure(s) public speak out their opinions towards policy and make adjustment to the proposals.'
	P2Q1L2.4: 'As public's voice (has) been heard, the policy carried out will run more smoothly and effective and cause less problem.'
	P2Q1L2.5: 'And so, amendment is no need after the policy has (been) carried out.'
	P2Q1L2.6: 'The government can have less staff to handle the following problem of polices and have between division of labour to other new tasks. Thus, effectiveness of governance is enhanced because of high degree of press freedom.'

Table AIII-4e Excerpts of *Formulation of Viewpoints, Opinions and Suggestions* from Paper 2 Question 1b in Samples E (Level 1)

Level 1 (Sample E from Candidate E)	
	Excerpt (from Paper 1 Q1(c))
Synthesise	*P1Q1L1.1:'I will suggest the measure which is provision of training to the newly arrivals from rural areas… In the retailing industries, the newly arrivals may learn about terms and skills with different kinds of customers.'*
	Excerpt (from Paper 1 Q2(b))
Synthesise	*P1Q2L1.1: 'The government in Hong Kong should ban 'medically unnecessary' plastic surgery it is because it might have an adverse impact on the youngsters personal development and may trigger health risks during the medical treatments of the plastic surgery.'*
	Excerpt (from Paper 1 Q3(b))
Synthesise	*P1Q3L1.1: 'It araise (arose) the world concern that once going to overseas, tourism should respect local cultures and have good manner or else it'll create dissatisfaction by the local people.'*
	Excerpt (from Paper 2 Q1(b))
Synthesise	*P2Q1L1.1: 'In Source A, it say the more press freedom a society has, less corruption, more efficient administration, higher political stability…'*
	P2Q1L1.2: 'The higher degree of press freedom can benefit the economic, social and political. All of the benefit can increase the credibility of the government.'
	P2Q1L1.3: 'The citizen(s) know they have a press freedom and they know the government will like to listen to their view, less argue(ment) that the citizen and the government will have.'
	P2Q1L1.4: 'The government can force on other problem and solve the problem.'
Counter-argument	*P2Q1L1.5: 'Some of the people may say their view will make a lot of argue(ments) when there is a high degree of press freedom.'*
Rebuttal	*P2Q1L1.6: 'But I think people will look for the compromise. The compromise can help the government to solve problem.'*
	P2Q1L1.7: 'It is a good choice to increased (increase) the credibility of the government since the government can provide a discussion for different view people and let them compromise.'

Appendix IV

Think-aloud Protocols

Table AIV-1: Think-aloud Protocol of Case 1

Date: 5/8/2015	**Time: 2:30-5:00**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 4) Taken the exam in English (Think aloud in English)
1	describe the trends in Source A
2	obviously an increase
3	1990 was 434
4	2012 was 1035
5	with a gradual change
6	down there, the trend of the profits matches with the one above
7	global problem
8	firstly more and more people
9	that means the problem
10	reason for the occurrence is related to the increase in no. of people
11	back to Source B
12	as 2005 to 2035 shows a progressive change
13	so I assume that the tourist arrivals from 2005 to 2035 keep increasing
14	obviously source B shows carbon dioxide emission
15	obviously related to environmental issues
16	obviously one of the global concerns is global warming
17	as it affects people in the whole world
18	we can see the increase
19	projecting to 2035
20	the increase was very big
21	and we can see that a great proportion of the carbon dioxide emission is related to transport
22	related to air transport
23	accommodation is related to tourism
24	tourists need to take airplanes
25	especially international tourists
26	transport is the major reason
27	to explain the global concern arising from the increasing trend in international tourism
28	the 2nd is from source C
29	Source C is about
30	it is about the Chinese
31	a guidebook for the Chinese to tell them to behave in a civilized manner
32	then
33	it also about cultural conflicts
34	people in different places have their own culture
35	for example
36	Source C mentioned
37	some mainland tourists urinated into a plastic bag outside the Golden Bauhinia Square
38	they don't find it a problem
39	but in fact
40	it may be offensive in other countries
41	because of the cultural difference
42	take Egypt as an example
43	carved names in temples
44	destroyed or harmed their own interests
45	yes
46	the 2nd global concern may be the protection
47	or conservation of the tourist spots
48	or heritage
49	in each country, the destruction or contamination of historical stuff

50	cannot be restored
51	i.e. difficult to be repaired or resolved
52	every country is concerned
53	because the quality and quantity of the heritage and the environment
54	whether visiting a country depends on the quantity
55	and the conditions of the heritage
56	for example, if I go to Egypt
57	if the Pyramids are destroyed, they are destroyed
58	cannot be recovered
59	use some examples to illustrate why
60	one of the global concerns is the countries
61	is worrying…
62	whether their heritage will be destroyed
63	they may not want people to come
64	because they don't want their properties to be destroyed
65	but at the same time they need to
66	i.e. they need to adapt to an international trend
	(when you see the question, what did you think of before you start answering your question? You mentioned global problems at the very beginning. How did you arrive at this?)
67	in the question, with reference to the sources is not important
68	identify and explain two global concerns from the trends is important
69	the trend is the one described in part (a)
70	correlate the sources
71	in a question, you need to make full use of the sources
72	useless information will rarely be provided
73	even if not reading the question
74	just reading the 3 sources
75	roughly you will know
76	you have got it in your mind
77	there is correlation between 2 sources
78	correlation between the trends in the sources
79	the following, need not mention it
80	many HK people are discussing about it
81	the interesting things
82	the first thing that comes to my mind
83	global concern is about magnifying some events in our daily lives
84	the mainlanders may be noisy and impolite in HK
85	they urinate in a plastic bag
86	in a global scale
87	they have their own culture
88	they have to adapt to another culture when in another country
89	obviously, there is a cultural conflict
	(Cultural Conflict is not found in the source. How did you arrive at this?)
90	Firstly, we have to find out why they did that
91	e.g. the source mentioned
92	for example, reminded the mainland travellers not to use their left hands to touch the others in India
93	as the Indians think that left hands are dirty
94	this is the tradition in India
95	a culture there
96	need to make a comparison
97	for the Chinese
98	using culture as the common…
99	to interpret the question
100	concerns are usually negative
101	not to touch them as it offends them
102	or they find it an inappropriate behaviour
103	so I used cultural conflict
	(besides from the source, how did your understanding of cultural conflict come about?)

104	when I said Indians don't want to be touched by left hands as they think that left hands are dirty
105	this is my knowledge
106	it's from the TV
	(Did you learn cultural conflict from lessons in school?)
107	no idea
108	I read a book in school
109	but there is no cultural conflict in the notes or powerpoints
110	such a simple word was not used to explain this simple stuff
111	they used heterogeneity
112	I don't know whether I am correct with this word
113	they used some difficult terms
114	in the exam
115	even though I am sure that this is the word
116	I won't use it
117	we need to make it simple and easy to understand
118	for the marker to understand my ideas
	(Why did you emphasize the use of examples in explaining cultural conflicts? Did you use examples because they are found in the source?
119	very common
120	when you say that this is cultural conflict, it is a statement
121	to justify this statement
122	or to make people believe what you say
123	you have to use some examples from reality
124	in LS, you must use some incidents or objects to support your statement
	(so you did it automatically)
125	identify refers to the statement
126	it says explain
127	to explain means you need to support your statement

Table AIV-2: Think-aloud Protocol of Case 2:

Date: 5/8/2015	**Time: 2:30-5:00**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 3) Taken the exam in Chinese (Think aloud in Cantonese)
1	the global concern is
2	the trend was the increase in arrivals by two-folds
3	therefore…
4	when there are more people
5	but no…
6	need to refer to Source B and Source C
7	the most obvious increase is found in transport
8	the whole source C is about cultural conflicts
9	i.e. when asked about global concern, I think these are correct
10	how to relate to the tourist arrivals
11	because there are more people
12	and then…
13	resulting in these two global concerns
	(when answering the question, what did you do first?)
14	because there are more people, there is air pollution
15	because there is CO2 emission in Source B
16	that is one point
	(How did you relate the increase in no. of people and the problem?)
17	it is found from source B
18	When I see CO2 emission, I feel that it is related to air pollution

19	maybe related to tutorial schools
20	Source C is about the cultural difference between the mainland and the foreign countries
21	I treat it as a global concern
22	and then…
23	cultural conflicts
24	I have not much to answer
	(How did you get to cultural conflict?)
25	I find that it is about civilized manners…referring to mainlanders
26	and then it says in the following
27	tourists have their own set of standards
28	it should be about the different ways related to different cultures
29	it says among HongKongers and Mainlanders…
30	that means what…
31	because of growing dissatisfaction means conflicts
32	taught by tutorial schools
33	I treated it as a comprehension
34	made a summary in my answer
35	the sources are the answers because it says with reference to the sources
	(What in the question triggered you to search for points of reference in the sources?)
36	global concern
37	it means something negative
38	even before taking a look at the question, when reading the sources, I guess it is about something negative
	(When you answer the question, did you think of what you have learnt in the tutorial classes? What were drilled or taught in tutorial classes?)
39	e.g. cultural conflicts
40	I have forgotten everything
41	I memorized a lot of stuff at that time
42	In the exam, I answered with all these memorized stuff
43	even in part (a), which is about potential benefits, I thought of air pollution right the way
44	I have got a lot of stuff learnt from tutorial classes in my mind. When I see a key word, it will put down the related part as answers
	(Is the stuff you learnt from tutorial classes categorized in a certain way?)
45	In the tutorial classes, we have got a lot of standard answers to different questions
46	these are arguments
47	e.g. the effects and problems of air pollution
48	The questions for sure will be found in the exams
49	When first taking a look at this paper, I was quite happy as the questions are all as expected

Table AIV-3: Think-aloud Protocol of Case 3

Date: 5/8/2015	**Time: 2:30-5:00**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 4) Taken the exam in English (Think aloud in mixed code (Cantonese + English))
1	describe the trends in Source A
2	obviously an increase
3	1990 was 434
4	2012 was 1035
5	with a gradual change
6	down there, the trend of the profits matches with the one above
7	global problem
8	firstly more and more people
9	that means the problem
10	reason for the occurrence is related to the increase in no. of people
11	back to Source B
12	as 2005 to 2035 shows a progressive change
13	so I assume that the tourist arrivals from 2005 to 2035 keep increasing
14	obviously source B shows carbon dioxide emission
15	obviously related to environmental issues
16	obviously one of the global concerns is global warming
17	as it affects people in the whole world
18	we can see the increase
19	projecting to 2035
20	the increase was very big
21	and we can see that a great proportion of the carbon dioxide emission is related to transport
22	related to air transport
23	accommodation is related to tourism
24	tourists need to take airplanes
25	especially international tourists
26	transport is the major reason
27	to explain the global concern arising from the increasing trend in international tourism
28	the 2nd is from source C
29	Source C is about
30	it is about the Chinese
31	a guidebook for the Chinese to tell them to behave in a civilized manner
32	then
33	it also about cultural conflicts
34	people in different places have their own culture
35	for example
36	Source C mentioned
37	some mainland tourists urinated into a plastic bag outside the Golden Bauhinia Square
38	they don't find it a problem
39	but in fact
40	it may be offensive in other countries
41	because of the cultural difference
42	take Egypt as an example
43	carved names in temples
44	destroyed or harmed their own interests
45	yes
46	the 2nd global concern may be the protection
47	or conservation of the tourist spots
48	or heritage
49	in each country, the destruction or contamination of historical stuff
50	cannot be restored
51	i.e. difficult to be repaired or resolved

52	every country is concerned
53	because the quality and quantity of the heritage and the environment
54	whether visiting a country depends on the quantity
55	and the conditions of the heritage
56	for example, if I go to Egypt
57	if the Pyramids are destroyed, they are destroyed
58	cannot be recovered
59	use some examples to illustrate why
60	one of the global concerns is the countries
61	is worrying…
62	whether their heritage will be destroyed
63	they may not want people to come
64	because they don't want their properties to be destroyed
65	but at the same time they need to
66	i.e. they need to adapt to an international trend
	(when you see the question, what did you think of before you start answering your question? You mentioned global problems at the very beginning. How did you arrive at this?)
67	in the question, with reference to the sources is not important
68	identify and explain two global concerns from the trends is important
69	the trend is the one described in part (a)
70	correlate the sources
71	in a question, you need to make full use of the sources
72	useless information will rarely be provided
73	even if not reading the question
74	just reading the 3 sources
75	roughly you will know
76	you have got it in your mind
77	there is correlation between 2 sources
78	correlation between the trends in the sources
79	the following, need not mention it
80	many HK people are discussing about it
81	the interesting things
82	the first thing that comes to my mind
83	global concern is about magnifying some events in our daily lives
84	the mainlanders may be noisy and impolite in HK
85	they urinate in a plastic bag
86	in a global scale
87	they have their own culture
88	they have to adapt to another culture when in another country
89	obviously, there is a cultural conflict
	(Cultural Conflict is not found in the source. How did you arrive at this?)
90	Firstly, we have to find out why they did that
91	e.g. the source mentioned
92	for example, reminded the mainland travellers not to use their left hands to touch the others in India
93	as the Indians think that left hands are dirty
94	this is the tradition in India
95	a culture there
96	need to make a comparison
97	for the Chinese
98	using culture as the common…
99	to interpret the question
100	concerns are usually negative
101	not to touch them as it offends them
102	or they find it an inappropriate behaviour
103	so I used cultural conflict
	(besides from the source, how did your understanding of cultural conflict come about?)
104	when I said Indians don't want to be touched by left hands as they think that left hands are dirty
105	this is my knowledge

106	it's from the TV
	(Did you learn cultural conflict from lessons in school?)
107	no idea
108	I read a book in school
109	but there is no cultural conflict in the notes or powerpoints
110	such a simple word was not used to explain this simple stuff
111	they used heterogeneity
112	I don't know whether I am correct with this word
113	they used some difficult terms
114	in the exam
115	even though I am sure that this is the word
116	I won't use it
117	we need to make it simple and easy to understand
118	for the marker to understand my ideas
	(Why did you emphasize the use of examples in explaining cultural conflicts? Did you use examples because they are found in the source?
119	very common
120	when you say that this is cultural conflict, it is a statement
121	to justify this statement
122	or to make people believe what you say
123	you have to use some examples from reality
124	in LS, you must use some incidents or objects to support your statement
	(so you did it automatically)
125	identify refers to the statement
126	it says explain
127	to explain means you need to support your statement

Table AIV-4: Think-aloud Protocol of Case 4:

Date: 9/10/2015	**Time: 11:55-12:30**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 5) Taken the exam in English (Think aloud in mixed code (Cantonese + English)
1	global concern --this word made me think of global warming etc
2	globalization problem
3	directly related to the graph
4	the major item in this graph is air transport
5	the predicted in 2035 increases greatly from 05
6	car transport drops a lot
7	the underlying meaning is that car transport represents short distance travel & air transport is for long distance travel
8	this represents as time goes by
9	the average travelling distance of tourists is very long
10	the problem brought about is air transport emits a lot of CO2
11	aggravating global warming
12	as one of the global concerns
13	…
	(What are you looking at? What are you thinking of?)
14	when I first answered this question, I missed the last part of the question and now I looked back to it
15	need to match it
16	coz when answering LS questions, we have to be accurate in answering each word
17	(What did you miss just now?)
18	when I first answered this question, I missed the last part of the question and now I looked back to it
	(What are you thinking about?)
19	the previous point
20	besides comparing the two years, the total has increased a lot
21	matching what is found in A
22	more and more tourists
23	the other one should also be discussed from this
24	more international tourists
25	the flow of tourists in the world increased
26	the other thing that can be found
27	the international receipts increased
28	the proportional increase is greater than that for the number of tourists
29	the problem brought about is
30	should put it like this
31	the receipts from tourists increased
32	referring to Source C
33	Source C is mainly about cultural conflicts
34	when tourists go from one country to another, if their culture is different, conflicts occur
35	the problem will be more serious
36	no. of tourists increased, from A
37	secondly, the receipts from tourism increased
38	this makes people think of the give and take
39	to develop tourism further even with greater cultural conflicts
40	to explain the cultural conflict further
41	to use the examples in the Source
42	like, like
43	peeing in the Golden Bauhinia Square
44	some mainland tourists visited the Golden Bauhinia Square
45	a mother helped a child to pee into a plastic bag in public places
46	maybe this is common on the mainland
47	those in China may think that there is no problem with it
48	in Hong Kong people may think that this should be done in the toilet

49	it is not appropriate to do it in public places
50	and therefore…
51	as time goes by, there are more and more problems like this
52	not only these, for example, spitting
53	Hong Kong people will be discontented with the mainlanders
	(Are you going back to the question?)
54	just explained one of the examples
55	going back to the question
56	it is about global concerns
57	one of them, relating to it, is cultural conflicts
58	from this example, we can find that tourist travelling to different places
59	Source A shows that receipts from tourism may increase in some countries
60	some countries, which urgently need the receipts from tourism
61	there may be impact of foreign cultures on local culture
62	in the long run, it may cause erosion of local culture
63	culturally, there is a tendency of monoculture in the world
64	(Are you going back to the question again?)
65	in fact, I think that those parts of my answer in the middle
66	the causes and influences are not related together well
	(What you said just now is a initial line of thinking? If you are really about to write it down, what will you do?)
67	how to relate them?
	What is the problem with your previous answer?)
68	it is not sound…
69	it is not sound when it goes from cultural conflicts to monoculture
70	…
	(What ate you thinking about?)
71	…
72	nothing to be added
73	If I am really in the exam, I will write the first point first
74	I will think further how to improve the second while writing
75	…
	(How will you improve it?)
76	will relate it more closely to the word global
	(In which of previous part is global missed out?)
77	just now I provided an example of an individual place
78	I did not explain how cultural conflict is happening globally
79	maybe one or two sentences can be added
80	as said before, there are more long distance travels
81	they visit different parts of the world
82	this increases similar cases of cultural conflicts in different places
	(Just now there were moments of silence. Did you try to link up your answer with your lessons? How did you come up with global warming & cultural conflicts?)
83	…exam skills in the lesson
84	every word in a question is useful. Need to make clear the meaning
85	make full use of all the sources
86	if the definition of some terms are useful, I need to recite it
87	But this is not needed for this question

Table AIV-5: Think-aloud Protocol of Case 5:

Date: 9/10/2015	**Time: 2:45-3:30**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 4) Taken the exam in English (Think aloud in English)
1	Source B is a graph
2	we have two years and emissions from different sources
3	accommodation and air transport account for the greatest amount
4	and then for…
5	Source C echoes
6	the data about the behaviour of tourists
7	creating dissatisfaction because of their misbehaviours
8	and then…
9	think of a global concern from Source B as environmental concern
10	Source C…I think it's about the bad tourist manners or their behaviour
11	they are causing disturbance to the locals
	(What are you thinking about now?)
12	When I answer the question, I will mention what the two global concerns are
13	environmental concern and behaviour of tourists
14	and then
15	when I talk about the environmental concern, I will describe the increase in the emissions of CO2
16	from 2005 to 2035
17	I will also talk about the components of emissions
18	the majority of the emissions is from air transport
19	while the least is from other transport
20	and then…
21	we need to relate Source A
22	we can see that the international tourism is expanding
23	and then…
24	it is obvious that air transport will thus increase
25	it is causing the global concern of environmental deterioration
26	the increase of international tourism
27	and for Source C
28	relating Source C with Source A
29	I will talk about the increase in international tourism
30	and then there are more foreign tourists going to different countries
31	and then they may have conflicts of culture
32	therefore may create dissatisfaction
33	or some misunderstanding
34	because of the difference in cultures
35	and standards of behaviour
36	and then I will make a short sentence about cultures
37	…
38	I will remind myself to make use some examples from the sources
	(What makes you think of using examples?)
39	my teachers usually remind us to make use of examples
40	I personally think that I need to make use of examples from the sources
41	or possibly some real life examples
42	though I don't think that it is possible for this question
	(What are the examples that you are going to use for this one?)
43	I need to think about it
44	let me think
45	maybe I will talk about the behaviour of the mainland tourists
46	e.g. they are jumping the queue
47	and creating discontent of Hong Kong people
48	and…
49	when quoting from the sources

50	I'll talk about the major emissions
51	using Source A, I will talk about the increase in international tourist arrivals
52	it rises from 434 million in 1990 to 1035 million in 2012
53	there is a huge increase in international tourism
54	I will conclude it in that way
55	…
56	I will also look at the key words, identify and explain
57	identify tells me that I should point out some trends maybe
58	or
59	point out some…here is talking about concerns
60	I think I will…
61	identify is more like point out
62	and explain…it's talking about reasons or the causes or phenomenon
63	in this case, it's talking about how the global concern is actually related to international tourism
64	it mentioned that "you described at (a)" and so I will look at (a) again
65	of course, just a brief look
66	and the words "with reference to" makes me think of adding more examples
67	I was taught that "according to" and "with reference to" have different meanings
68	and so I'll look at these words as well
69	and also the number 2
70	it's focused. Also, it's highlighted
71	…
72	I am thinking about what I should actually do
73	actually I remember when I answered this question, I did not have much time left
74	and this was also my worst question
75	I am looking at what else I can do when answering this question now
76	I guess the source is talking about different countries
77	India… and then this is Hong Kong
78	as it's talking about international tourism
79	maybe India is also a good example to be quoted
80	if I have time
81	a mother helping her son to urinate into a plastic bag is also an example
	(What will you do with the examples?)
82	I will choose one or two from each source
83	For this one, it's about data
84	and this one… there are some examples
85	I will try to integrate one or two into my arguments
	(What's the argument that you want to make?)
86	The second global concern is misbehaviour or the conflicts of culture
87	I guess conflicts of culture may be my final choice
	(Why did you choose it?)
88	because global concern
89	because misbehaviour is like an action
90	I mean…
91	it's something that is causing discontent
92	so that should be conflicts
93	it's more like a global concern
94	I'll talk about conflicts of culture instead of misbehaviour of tourists
95	…
96	usually I read the questions before I read the source in the real situation
97	because first I'll look at the key words to see what I have to do
98	for this graph, I don't have much to talk about
99	It's just about increasing CO2 emissions
100	I will just point out the majority
101	…
102	I don't think I can add anything from this graph
103	just an increasing trend
104	for Source A, I may wonder whether I need to include international tourism receipts

105	it's also an evidence about international tourism
106	but I don't know whether I should put down only international tourist arrivals or both of them
107	I guess it just depends on my time
108	if I have more time, I may add one more
109	I may also talk about international tourism receipts will increase from 262 million in 1990 to 1078 billion in 2012
110	although I think that these two are talking about the same thing
111	international tourism is actually expanding or increasing
112	yes, that's it
	(Why do you think that one of the concerns is deterioration of the environment?)
113	I think that it is an environmental problem
114	deterioration of environmental quality is an elaboration
	(How did you get it from the graph?)
115	carbon dioxide is greenhouse gas
116	increasing carbon dioxide will enhance global warming effect
117	then air pollution increases
118	causing environmental problems in the end
	(Why do you think that these are related to the question?)
119	Global concern is like a problem that should be solved
120	environmental problem is a problem
121	if I just say carbon dioxide increased, it is not yet a problem
122	it's just a phenomenon
123	and then…
124	need to convert to an environmental problem for it to be a problem
	(Don't you think that there is more to be elaborated on why global warming is a global concern?)
125	If there is a separate question on that, I will.
126	But there are only around 10 minutes for each question, I won't be able to put in such details
	(Don't you think that there is more to be elaborated on the link between cultural conflict and global concern? Don't you take them to be the same?)
127	I don't think they are the same. But when answering the question, I will just put down the second concern is conflict of cultural
128	if you ask me now, I think conflict causes disputes between local citizens and tourists
129	Even if there is tourism, it is disharmony
	(Have you thought of values when answering the question?)
128	values…
129	No
	(Are there much discussion on beliefs or values?)
130	I remember that… we have discussed values in related to the political (aspect)
131	we have learnt about "global"
132	If we see "global" in the question, usually we need to use more examples
133	There are more in lessons on political modules, e.g. freedom, or what is it called?... Democracy
134	not much in the environmental modules
135	we have learnt the conflict between environmental protection and economic development
	(So you have not considered why people are so concerned about cultural heritage and you have not considered the values behind the cultural conflict. Is it because it is less discussed?)
136	my teachers in S4, 5 and S6 were different
137	The teacher in S4 & 5 was inexperienced in teaching LS
138	The one in S6 taught us to use some key words related to LS
139	not values
140	She analysed the questions
141	The one in S4 &5 taught us cultural heritage
	(Did you make use of exam skills?)
142	yes, e.g. better to have some authentic examples
143	read the sources carefully
144	and time management
	(Did you relate to your memory of the key words, e.g. cultural heritage, when you answer this question?)
145	For the first one, I identified something related to the environment right the way
146	for the second, I thought it was related to manner & misbehaviour
147	and then I thought conflict of culture is better

148	I could not identify that it is related to heritage

Table AIV-6: Think-aloud Protocol of Case 6:

Date: 9/10/2015	**Time: 3:35-4:15**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 5*) Taken the exam in Chinese (Think aloud in Cantonese)
1	I first find out what type of question it is
2	this is "explain", with analysis
3	it says "with reference to", then I take a look at the hints in the source
4	this one is about culture
5	C is about different places have different cultures
6	when going to other places, there will be conflicts
7	this may cause conflicts among the locals and tourists
8	…
9	now I am looking at B
10	in B, looking for the items accounting for the largest proportion in the chart
	(Which item?)
11	air transport
12	and take a look at the changes between year 05 and year 35
	(Which part are you reading or what are you thinking of?)
13	finding out the change
14	air transport increased by 10%
15	but vehicles decreased
16	as a whole, it was about 1200 in 05, but it was nearly 3000 in 35
17	the difference is a multiple of 2 point something
18	and then I will think of how to answer
19	now I will construct it with concepts
20	for this one, the global concern, I would say, is the environmental problem brought by tourists
21	for C, I would say it is about social
22	the I will start to see how the sources can be used
23	and see whether there are actually real life examples
24	for this one, B, mainly quoting the figures
25	make the most of the figures
26	and the change in terms of the number of times of increase
27	in general, the behavioural changes
28	quoting some examples, like talking loudly, spitting
29	…quote some behaviours of the tourists in different places
30	e.g. in Rome. Some tourists made some graffiti in the Colosseum
31	even these are not found in the source, it is related
32	this is included to let marker know that I am not just copying from the source. But reading news often
	(What will you do afterwards? Don't you think that you have answered the question?)
33	this is roughly the framework
	(How will you make use of the examples to write about the two concerns?)
34	like playing with building blocks
35	topic sentences are found in the first and the last sentence
36	the first concern is the environmental problem
37	use the data
38	and then explain the possible environmental problems brought by CO2
39	bringing about global warming
40	or extreme climate
41	the increase in CO2 is not yet relating to the global concern
42	more CO2 is the fact
43	the concern maybe about the problems brought about

44	and then explain
45	after explaining, the paragraph will end and I will open a new paragraph
46	the new paragraph is about the social aspect
47	there will be conflicts between tourists and the locals
48	quoting some examples from Source C
49	and then add some examples of my own
50	and then explain how these conflicts arouse concerns
51	affecting social harmony
52	deterring tourists
	(Why did you use examples? Why did you explain that global warming is a global concern?)
53	if there are no examples, there is no ground
54	it says "with reference to" and it's given to you for your use
55	This is my way of doing it
	(How did you go from conflicts to harmony?)
56	I was thinking about Hong Kong
57	in Hong Kong, there are travellers with parallel goods
58	it's something similar
59	mainland tourists coming to Hong Kong behave in manner different from Hong Kong people
60	and then they (Hong Kong people) may have some reactions, and conflicts
61	as I have seen, it affects social harmony
62	that's why I add this to the answer
	(Why do you think that harmony should be mentioned instead of stopping at the examples?)
63	when it comes to concerns, they may not be concerned about the conflicts
64	maybe they are concerned about harmony, which is of a larger scope
65	…
66	finished
	(Is there anything you have learnt in LS that helps you come up with harmony in your answer?)
67	school teaches us to deduce
	(How to deduce?)
68	e.g. what will be the problems caused by global warming
69	or what happens with conflicts, disputes?
70	going one or two steps further to wrap it up in a better way
	(How did you deduce? Did you think of some related concepts when you go from conflicts to harmony?)
71	I just think about what happens afterwards if there are disputes
72	when I think of a sentence, I will think about the concepts that I may use to wrap it up
	(Is there much discussion of harmony, conflicts or much broader, values?)
74	Not too much
75	No specific discussion on these
	(But discussions of harmony and conflicts are discussions of values.)
76	Maybe I did not realise that these are values
77	it might be immersed
78	e.g. when discussing Occupy Central earlier, or demonstrations
	(Did teachers take any stance in these incidents
79	No, they did not take any stance
80	At the beginning, facts were introduced
81	and then gave some broad questions for us to discuss
82	e.g. Should human take priority over rule of law?
83	Don't you build up your own standpoint in the discussions
84	yes, I have got my own stance
85	try to explain my own arguments
86	it's just like answering questions
87	sometimes, there are questions about your standpoint
88	we use evidence to support our arguments
89	we may use examples or data to support
90	teachers, in the end, will explain the possible arguments to support viewpoints on different sides
	(In the discussions, can you tolerate people with views different from yours?)
91	I don't have a strong intention to rebuke
92	As long as the reasons he put forward is logical, rather than just to oppose/ pinpoint somebody

93	as long as he is able to put forward a standpoint with supporting arguments, I think I am OK
	(What do you think of the relationship between human and the environment? Is it in harmony or is the environment just something for exploitation?)
94	I'm more on the side of harmony
95	If it is within my capacity, I will be more environmentally friendly
96	Sometimes, we cannot be overly environmentally friendly
97	like not switching on the air conditioner even if the temperature is over 30 deg
98	We will use it if we should
99	if the temperature is only 20 something, then let it be, switch on the fan
100	if the homework is not that important, I will hand it in by recycled paper
	(What are the relationships with people? A win-win situation or competitive?)
101	win-win format
102	you don't need to be the top to control everything
103	even if you are not the top, you may voice your view or help and accomplish a task
104	it dosen't need to…
105	even if I am given the top position, I don't think I would be able to lead so many people or control them
	(When answering the question, did you consider environmental values/ the environmental perspective?)
106	I didn't…
107	I think that there is a relationship between environmental protection and the question
108	But I didn't think of using the environmental perspective to answer this question
109	I know that it is about the pollution/ damage brought about by tourism. But I didn't think of talking about environmental protection
110	This question is about concern. I understand it as the problem
111	like the problem brought about by tourism
112	and then this problem leads to concerns
113	this term is a bit vague, I think
114	Our school did not emphasize the reasons why cultural heritage is important
115	Why we need to concern about a certain problem
116	why the society needs to provide resources for the elderly and the poor
117	There is not much about these
	(Did you seldom discuss why a certain problem is important to us?)
118	Seldom
119	Often, we will be told of the background
120	then, what are the solutions to tackle this problem
121	and what may be the problems brought about by the solutions
122	maybe, we discussed the pros and cons of every solution
123	however, back to the basic…why we have to consider this problem
124	or what may be the impact of this problem on the society
125	or it may not be a problem
126	maybe why we need to protect the environment
127	What benefits can be brought by environmental protection to our future
128	there is not much discussion
129	when it comes to environmental protection, the discussion stayed at the perspective of…the problems
130	alleviating extreme climate or slowing down the rise in sea level
131	still staying at the level of problems
132	we have not discussed why we need to protect the environment
133	we should not over-exploit it
134	we need to reserve resources for the next generation
135	teach the next generation to use the amount that we need
136	there is not much discussion of this sort, values... in our school
	(But it seems that you have got some discussion on freedom, harmony, which are values.)
137	yes. Like when discussing Occupy Central, we discussed whether law and order or human rights should take priority.
138	But we did not have discussion topics like how important law and order/ human rights is
139	he (teachers) may think that the importance of something is what you should know
140	there is no need to teach it in class and we will be able to explain them
141	it might because schools did not discuss the importance of it, we cannot deduce/ discuss further
142	Maybe my school has done so. But I was not aware of it

Table AIV-7: Think-aloud Protocol of Case 7:

Date: 9/10/2015	**Time: 6:40-7:00**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 5) Taken the exam in English (Think aloud in mixed code (Cantonese + English))**
1	taking a look at the headings in Source B
2	found one of the global concerns in Source B…
3	CO2 emissions
4	and then read Source C
5	relating the sources with the question
6	finding out the two global concerns from the sources
7	and then now I try to confirm whether I can use these two
	(How to confirm it?)
8	maybe, finding out the key words, global concerns
9	and then I should not look at it regionally
10	in Source C, there is a destruction of tourist spots
11	for example, maybe the natural scenery
12	I need to be careful not to take the regional perspective
13	adopt a global sense
14	the problem happens globally
15	and then… the first problem is…
16	a concern about carbon dioxide emission
17	it's because…
18	extracting some evidence from here
19	for example, there is a lot with air transport
20	many tourists take the plane, contributing to air transport
21	a lot of carbon dioxide is emitted from air transport
22	and so international tourism is increasing
23	we have already found out from (a) that there is an increasing number of international tourists
24	firstly, the trend will be stated
25	and then the trend will cause an increase in the emission of carbon dioxide
26	and then extract the data here as a support
27	it increased from 43% to 53%
28	and then B (C) is about the destruction to the natural scenes of countries
29	or the destruction of natural scenes and tourist spots of countries
30	the paragraphs in Source C are providing supporting arguments and examples
31	for example, some Chinese tourists carved their names in an ancient temple
32	this is the framework of the answer
	(Then, how would you further structure your answer and use the sources to answer to question in detail?)
33	structured into 3 paragraphs
34	I will sketch it first, but not in detail
35	I will write down the trend
36	increasing number of tourists
37	and then two global concerns
38	the first one maybe increasing amount of carbon dioxide
39	C is about the destruction of constructions
40	and then…
41	yes, this is the main structure
42	I will not think in detail
43	and then think of how to support the arguments while writing out the answer
	(If you are asked to write out the answer, how will you continue to answer the question?)
44	citing Source B
45	cite the trend, the key points
46	with reference to Source B, there is an increase in the emission of carbon dioxide by air transport
47	and then explain how it comes from air transport
48	as tourists are travelling by airplanes
49	and then air transport causes an increase in carbon dioxide emissions

50	if there is an increase in the number of tourists travelling
51	there will be an increase in carbon dioxide emissions
52	this is a global concern because it may cause some problems
53	carbon dioxide emissions may cause some problems
54	for example, global warming and pollution
55	the structure is more or less like this
56	Source C, first, about the destruction of tourist spots and constructions in some countries
57	and then…
58	because of the increase in the number of tourists travelling
59	there will be more travellers in different countries
60	it is difficult to regulate the behaviour of tourists
61	if we don't regulate, some tourists may destroy our constructions
62	for example,
63	with reference to Source C
64	using the carving of names in an ancient temple in Egypt as an example
65	this example shows that we cannot regulate the activities of tourists
66	and then an increase in the number of tourists will increase the possibility of such a destruction
67	in the last paragraph, because of the increasing number of tourists, the two concerns will arise
68	using Source C
69	this concern may lead to a problem of destruction of tourist spots in other countries
70	the others cannot enjoy them
71	this will affect the whole world
72	as these are international tourists, they may be from different countries
73	this will turn the problem global
74	owing to this, the locals may not be able to enjoy these
75	then this is a global concern
76	in the end, I will repeat the trend and the two global concerns
	(How did you relate the key word in the question, global concern, with the examples? What approach did you adopt in finding out the evidence? How did you relate destruction to the tourist spots and global concerns?)
77	something arousing problems is a concern usually
78	destruction of tourist spots will bring about some problems
79	and as it is found in the source, it is related to the question
80	we find that something is happening in Source C, which may lead to a problem
81	we may blow up a regional problem
82	to see whether the global situation is like this
83	if this is found globally, then it is a global concern
	(How did you explain that this is a global problem, not regional?)
84	There are a lot of international tourists
85	they will go to different countries
86	so the problem is not only found in a certain country
87	every country may have the chance to suffer from it
88	it's a problem in the global
89	bringing all countries in
	(Don't you think that global warming means a global concern and there is no need to explain further?)
90	yes, I'll elaborate on it
91	it is a concern because there is a problem
92	as global warming will lead to environmental problems
93	glacier in the North Pole will melt
94	as the whole world will face this problem, it will be a global concern
95	we can't say that a problem is a concern
	(Did you think that the tourists' behaviour is insulting? Have you thought of how people might be hurt or how national sentiment might be affected?)
96	Why carving a few words hurt/ offend you?
97	a few more lines have to be added to explain it
98	it's complicated
99	and so at the beginning…
100	it's hidden
	(Don't you think that it is a reasonable approach to think along the line of affection? In comparison with enjoyment, which was in your answer, do you think that affection is an important factor?))

101	I think that it is not the most important
102	nowadays, people's concerns about their nations are not high
103	The insult might not be felt by a lot of people
104	the scope of influence may not be large
	(Besides the sources, what did you relate to when answering the question?)
105	it might not be necessary for this question
106	But for some questions, maybe a platform will be used for comparison
107	I might need to think more about the examination skills
108	But probably not for this one
109	it does not make exam skills prominent
	(What do you mean by examination skills?)
110	how to answer a question
111	how to answer a question on effectiveness
	(What's the relationship between how you answered this question and what you have learnt in class?)
112	I have not specially recall what I have learnt in class
	(Did you answer this question with anything you have learnt in class, for example, concepts?)
113	I don't see the need in using concepts in this question
114	for example, some questions may be about quality of life
115	then we need to talk about concepts
116	but there is no need for this one
	(To you, does using concepts mean writing down the definition?)
117	When you really need to use quality of life…
118	it's about increasing your satisfaction towards material and non-material life
119	when you answer the question, you will think about whether an action will increase your satisfaction
120	the definition helps the subsequent parts of your essay
121	the subsequent parts of your essay has to be related to the definition to explain how quality of life is improved
	(Does it mean that you don't consider cultural clash and global warming concepts?
122	cultural clash…
123	cultural clash can be considered concepts
124	but I will not give a definition specifically to cultural clash
	(Did you think of concepts that are related to the question?)
125	Seldom
	(Usually you answer questions by referring to the sources, and then answer according to the key words of the question.)
	(Recalling your lessons, did you learn anything about values? For example, was there much discussion on cultural clash, environmental values?)
126	Not much
127	We seldom discuss a value
128	We usually discuss a case, with a given stance
129	Very often, we were pre-assigned a stance. So little value judgement of our own
	(So not much discussion of the value positions underpinning a certain stance.)
130	Seldom
131	Usually the discussion might be brought out. But we seldom discuss values on its own
	(What do you the relationship between human beings and the environment? In harmony? The environment is for your use?)
132	I think I will strive for a harmonious relationship
133	This harmonious relationship can be viewed from a practical perspective
134	If we continue to exploit it, it will exhaust one day
135	At that time, both will lose
136	we can't afford this situation
137	then we want to have a harmonious relationship
138	the second one is a natural bond with the environment
139	but we cannot feel it strong in the urban environment
140	but we don't want to exploit it
141	don't want to win over it
	(Which one is your line of thought?)
142	the practical one
	(Getting back to the answer, why did you emphasise the use of examples and evidence?)
143	On one hand, schools emphasized evidence

144	secondly, if there is no evidence, it seems ungrounded
145	you don't have concrete evidence for support…
146	maybe, just assumptions
147	not convincing
148	so there must be examples
149	examples help to make your elaborations reasonable

Table AIV-8: Think-aloud Protocol of Case 8:

Date: 22/10/2015	**Time: 2:15-2:45pm**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 5*) Taken the exam in Chinese (Studying Earth Science) (Think aloud in Cantonese)
1	with reference to the source
2	the trend in Part (a)
3	the no. of tourists increased
4	the receipts increased
5	how the figures multiplied was mentioned just now?
6	and then… the multiples
7	the number in the lower part changed by a higher multiple
8	i.e. it is not in direct proportion
9	not more people…
10	though more people come
11	the increase in receipts was greater
12	the expenditure per person should be higher
13	maybe they are richer or higher income…
14	the expenditure per person should be higher
15	this can be found
16	read it again
17	when the expenditure per person increase, what will be the concern?
18	then…read the figures later
19	the heading
20	usually read the heading first
21	From the heading, I should be able to guess that it was environmental protection
22	though I could guess it, of course I need to read it.
23	and then from the news…
24	skimmed through quickly
25	it's about civilization
26	of course, I need to read it carefully
27	and then… as it says "referring to the sources", I could write about other stuff (not in the sources)
28	besides those in the sources, if I came up with other examples, I could also put it down.
29	going back to the previous part
30	…
	(Which part are you reading now?)
31	Not yet read it. Very often, I will think of how to answer after reading
32	I usually read the headings and make a guess on the points
33	and then think carefully how to fit the data into it
34	it says two…roughly it is about these.
35	carbon dioxide emissions… Expected to rise
36	as usual…calculate it
37	though it is rising, the percentages (for the items) are different
38	but the difference is not great
39	the difference mainly lies in air transport
40	cars…
41	guess that air transport means more transnational (travel)

42	for the percentage of road transport, though it is longer, it is lower
43	there should be an increase
44	but the proportion in the total is smaller
45	there is an increase
46	that's all from air transport
47	air transport… aeroplanes is the most environmentally unfriendly means of transport
48	I know that it is about more and more people visited foreign countries…I think…
49	From this, air transport increases
50	and then… not much about the others
51	mainland visitors in Egypt
52	silence
53	…then… civilization
54	destroying…
55	silence
56	copying from the source, is it?
57	the first point is about the consideration of the environmental aspect
58	more pollution
59	and more emissions
60	causing the greenhouse effect
61	then about the concern over this
62	it says two global concerns
63	we need to explain why this is global
64	carbon dioxide can be a local pollutant
65	that is air pollution
66	need to talk about global concerns by referring to the effect in the world
67	e.g. not only in local
68	but in the world, e.g. rise in sea level and so on
69	people in the world will be affected
70	as so this is a global concern
71	then… yes, should answer like this roughly
72	this question… that is the second point is about civilization
73	civilisation…
74	the mainlanders may be reluctant in consumption when travelling in foreign countries
75	I've heard some news about mainlanders eating cup noodles, instead of spending outside
76	not enhancing the local economic development
77	they used flash lights to take photos in museums
78	maybe they do not follow the rules
79	or maybe they are less educated
80	besides mainlanders, some people in other countries may not respect the local culture when travelling abroad
81	or national…
82	maybe… when you are in Rome, do as the Romans do
83	girls may put on headscarfs round their heads when in the Middle East, even though they are foreigners
84	these are something related to civilization…
85	not about civilisation…
86	but just paying respect to the locals
87	as you are visiting that place, you need to do these
88	this may reflect…
89	but this is not really the concern
90	this…
91	civilisation…
92	the concerns should be respect for culture and regions and so on
93	maybe the destruction to cultural heritage
94	why should the world be concerned about it
95	maybe if tourism
96	if you make the locals unhappy
97	that is disrespect
98	they may not welcome you
99	maybe both parties will not be happy

100	don't know how to describe it
101	need to think about it
	(Along which direction are you thinking?)
102	think of some related terms
103	I have not used these for some time
104	it should be related to civilization and culture
105	also about these…
106	copy that in and change some words
107	in general, it's like this
	(In fact, you will think of what you have learnt in class, like some terms or concepts.)
108	very practical. Just fit them in
109	to show that I know something
	(Have you learnt global warming in class?)
110	this one… global warming this term came more naturally than the other
111	Maybe because I have studied Geography
112	I don't know. But I linked it up with a lot of things
113	Maybe I have got some revision on this part
114	it must be… some stuff in LS is also found in other subjects
115	I was not so familiar with the other concern
116	got stuck for a while
	(How did you think of the terms at that time?)
117	tried to see whether I have learnt some related…stuff
118	maybe tourism in China… on the mainland
119	something bad… like providing something artificial just to gain profits
120	but they may lose attraction to tourists and it will be difficult for them to make a living
121	but this is not related to civilization
122	civilization is mainly related to museums, toilets and something related to culture and ethnicity
123	the first thing I thought of was people in the Middle East
124	people with different cultures
125	Or when you visit India, as far as I know, they are not so concerned about the use of right or left hands. Not so outrageous
126	think of something about civilization, if I know it
127	If I think of it, I will write it down
128	it says "with reference to", I write it down if I can think of it
129	I will think of some examples while I am writing out the answer
130	I will not think of all the examples before writing out the answer
	(What kind of examples?)
131	elicit some examples from the source
	(Why did you put down some examples?)
132	because it says "with reference to the sources"
	(How did you use the examples?)
133	there may be news reports on these examples
134	news reports reflect the global concern
135	that means many people will be concerned about it
136	with a lot of reports and discussions
137	and then link it to the question
	(Any supplement to the answer?)
138	it just says two points, identify and explain
139	just identifying two points and then explain it
140	just think about how to explain
141	the question is clear… identify and explain
	(What comes to your mind when you see "explain" in the question?)
142	examples and evidence
143	evidence should be something that has happened… in the news
144	not something made up
145	and explain how the example is related to the trend
146	when more people are visiting there, the quiet environment will be destroyed
147	and then explain how this example is related to the concern

	(How to explain it?)
148	it is not natural
	(Why? Did you link up the example and the concern?)
149	when there are news reports on it, it means global concerns
150	if a single example is required, the question should be about the cultural training of mainland tourists
151	it is not about global concerns
152	if there are a lot of reports on it, it can prove that many people are concerned about it
153	if people are not concerned about it, there would not be so much discussion
	(Why do you think that evidence is needed in answering a question of explanation?)
154	that is how I understand the question
155	evidence is needed for answering a question of explanation
	(How did you learn it?)
156	basically, all subjects got this requirement
157	in fact, Chinese or composition is also like this
158	just logical deduction.. Though nothing wrong…as I study Science
159	teachers always think that examples are needed
160	usually I use examples to explain my points
	(What have you learnt in LS that helped you answered this question?)
161	I learnt global warming in LS also
162	it is impossible to answer this question exactly
163	at least, I need to know how to interpret graphs
164	e.g. it (the bars) might seem to be longer, but the percentages are quite the same
165	e.g. in Source A, it is not just about increase by 2 and 4 times. But the consumption per person should be worked out
166	this graph is tricky and it requires careful interpretation
167	this is a complicated graph
168	we are familiar with the interpretation of different types of graphs
169	it might be related to (what we have learnt)
170	in comparison with a question with 3 textual sources, this one is easier
171	every subject may be related (to LS)
	(What do you think of the relationship between the environment and you? In harmony or is the environment for your exploitation?)
172	I am inclined to the view of loving the environment
	(What about your relationship with other people?)
173	it depends on the situation
174	e.g. when I took the course here, we are in harmony. Quite a lot of discussions. We didn't aim at getting higher marks than the others
175	maybe people here know that we are doing well, or better put it as better than the others.
176	we gathered here because we were interested in
177	in a normal secondary school setting, there are incidents without harmony
178	it can't be described as competition
179	I don't know the others' view
180	I don't mind sharing some notes for quizzes, past papers
181	[opened a group for my class and uploaded some exercises there
182	my classmates are friendly
183	there was a large proportion of male classmates in the Science class and the situation is somewhat like here
184	some other classes did not share things maybe they thought that they had paid for it
185	I shared some notes made by myself with my classmates
186	it might be because of my homeroom teacher, who teaches Chinese and I was influenced by the Chinese culture
187	He has influenced me a lot
	(Did you learnt what you mentioned, respect, culture in LS?)
188	Not much in LS… Maybe in Personal Development
	(What about values?)
189	there might have been discussions on that in LS. But more in Chinese

Table AIV-9: Think-aloud Protocol of Case 9:

Date: 22/10/2015	**Time:3:00-3:40pm**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 4) Taken the exam in Chinese (Studying Science) (Think aloud in Cantonese)
	(Which part are you reading?)
1	reading the question
2	it says "with reference to the sources", that means all the 3 sources should be used
3	Part (a) used Source A only
4	Sources B & C may be used
5	I will read the sources first and then answer
6	the trend described in Part (a)… so I should refer to the trend above
7	and then explain the how the trend is related to the global concerns
8	silence
9	B is about carbon emissions
10	it increases a lot in 2030
11	but the percentages are quite the same
12	increases as a whole
13	increases by at least 2 folds
14	among them, the greatest is with air transport
15	because of tourism, people are going from one place to another, air transport, this industry is important
16	the trend in (a) is about… more and more people travelling
17	the growth of tourism is great
18	this trend leads to the increase in carbon emissions
19	this is about the growth in tourism leading to the increase in carbon emissions
20	C is about…
21	C is about different cultures in different places
22	the people…the locals may not accept what the tourists do
23	it might not be correct…maybe it is wrong
24	maybe there will be conflicts
25	it's like this…
	(What are you trying to find out from the question?)
26	it says two…B and C… one is this one (B) and the other is that one (C)
27	this one is about… more and more people are travelling abroad and so there may be a higher chance of these incidents
28	though we don't know whether there are more mainland tourists, C is mainly about mainland tourists
29	maybe because of the culture…
30	very often, they behave in an uncivilized manner
31	there maybe be conflicts with the locals,
32	dislike or unhappy
33	it may be because there are more and more… the rising trend of tourism…
34	more people visit or tour round the place
35	or maybe do something…
36	mainly use these examples of uncivilised behaviours, like spitting
37	the locals may not find these acceptable
38	this one…
39	in terms of universal values, these are not acceptable
40	and people should not do it
41	it is because of tourism, mainlanders do some inappropriate things
42	as says here at the bottom… should respect the others
43	but in fact, there are a lot on the mainland
44	…
45	when answering this question, I will do paragraphing
46	as it says two, I will write in 3 paragraphs
47	an introduction
48	I will talk about how the trend above would bring about two global concerns
49	it's about the increase in carbon emissions and then some elaboration

50	add some examples
51	start from C and then the relationship with (a)
52	it causes conflicts about civilization
53	and then I think … add some examples from the Source
54	at the end, a conclusion
55	I will answer in this way
	(For the first concern, will you just say carbon emissions increase and this the global concern?)
56	…
	(Just now you mentioned elaboration. Does it mean some figures?)
57	…
58	I start with the increase by 2 folds
59	…
	(Did you think of what you have learnt in LS when answering the question?)
60	the concept of globalization
61	people are getting closer
62	advancement in technology makes travelling from one place to another faster
63	information flow is fast
64	but do not know how to use these
	(You mentioned conflicts. Have you got discussion about conflicts in class?)
65	Yes. There are something similar
	(What sort of discussion? Not of the same topic?)
66	Yes, for sure we had
	(After describing the figures about carbon dioxide emission, do you think that you have explained the global concern already?)
67	it may cause global warming, which people in the world will be concerned about
68	the temperatures rise, our daily routines may be changed
69	the world will be changed and many people can feel it and so people will be concerned about it
70	When answering the question, I thought that this is the point already. (did not elaborate on it in the answer)
71	Maybe it is because I study Science
72	But the others may not understand it
73	I might neglect the gap in between
	(You mentioned universal values. What do you think universal values are?)
74	I think everyone living in the society, on the Earth, should have some unique behavioural standard
75	Every place has its own civilization and we need to respect it when we are in that place
76	we should not be like the mainlanders, who spit, which is unhygienic
77	the society attach much importance to hygiene
78	because of diseases, hygiene is important
79	and not doing anything that may affect the others
80	do what we should do
	(Have you learnt universal values in class?)
81	Yes
82	When discussing values, universal values were also discussed
	(Did you discuss values on its own or discuss it with reference to some cases or incidents?)
83	Both in school
84	we discussed briefly about some current affairs
85	Our teacher will tell us about the concepts involved in some current affairs
86	and why did these conflicts occur
87	how are resources being utilized in the world
88	and different examples
89	and then some worksheets related to the modules
90	not many examples
	(What do you think of your relationship with the environment? In harmony, co-existing or the environment is for your exploitation?)
91	Co-existing
	(What do you think of your relationship with other people?)
92	If we need to guard against the others, it will be difficult
93	sometimes, it is no big deal to be disadvantaged
94	unless it endangers me

	(When you went for the DSE, did you share some resources with the others?)
95	Yes, I did
96	We shared the notes from tutorial schools
97	we did the exercises together
98	if we did not understand something, we put our heads together
99	we gained from sharing the exercises
100	we gained more
	(Back to universal values. Was that taught by the teacher?)
101	my LS teacher was good
102	at the beginning, I jumped from step to step, leaving a big gap
103	I thought that it need not be explained
104	my teacher reminded me to fill the gaps in my arguments
	(Just now you mentioned that people should be ethical. Where did you learn it?)
105	teachers and my family always say so
	(How did you fill the gap in the arguments?)
106	my teacher asked me to look back to my answer to see why it goes from one to another and whether there is something in between
107	e.g. if you say that you caught a cold because you did not cover yourself with a blanket, is there any more to explain it
108	maybe you switched on the air conditioner

Table AIV-10: Think-aloud Protocol of Case 10:

Date: 22/10/2015	**Time: 4:00-4:40pm**
Venue: The HK Academy for Gifted Education	
Task: 2015 HKDSE LS Paper 1 Question 3(b)	
	Male (Level 4) Taken the exam in English (Studying Social Sciences) (Think aloud in mixed code (Cantonese + English))
1	firstly, I read the question of Part (a)
2	it's about Source A…focus on Source A
3	it's about the trends and so I answered the trends first
4	it increased
5	for example, the increase from 1990 to 2012… increase in the arrivals and receipts…describe both
6	and then I will think of a potential benefit
7	the potential benefit… international tourism caused cultural exchange
8	because of globalization… transport is more convenient
9	the living standard is higher and people can travel
10	as they travel, the quality of life is increased
11	and people can learn different local cultures
12	…
13	I have finished with (a)
14	I will read (b)
15	…
16	after reading the question, I will read the source
17	silence
18	(What are you reading?)
19	I am reading the question
20	I will highlight identify and explain
21	the topic sentence is about identify and then explain
22	silence
23	the first point… Source B is about a lot of carbon dioxide
24	it causes air pollution
25	Source C is about the destruction of cultural heritage of tourists
26	what I have identified is… air pollution problem
27	secondly…because of the low educational level, some people do not know the local cultures when travelling
28	local… i.e. should not destroy public properties

29	these two are the global concerns
30	and then explain… I will explain pollution first
31	pollution is because people travel more and take planes
32	when taking different means of transport when travelling, air pollution occurred
33	air pollution damage cultural heritage
34	like Tah Mahar in India is deteriorating because of air pollution
35	air pollution causes acid rain
36	and then cultural heritage is weathered
37	the other global concern is… people do not respect cultural heritage
38	for example, I will quote the source
39	for example, in Egypt, a person carved his name in a temple
40	I will then quote another example…
41	many people litter and stay overnight in the Great Wall
42	for explaining… I will use some examples I know to explain and enrich my answer
43	…
	(Why did you use examples?)
44	I need to show to the examiners that I know this issue
45	i.e. we should not just quote the source
46	it's just like reading comprehension
47	I can make the answer more complete by giving examples
48	to show off that I have more knowledge of it
	(What have you learnt in LS that helped you answer this question?)
49	I remembered that concepts is emphasized in the assessment
50	terms like cultural heritage should be used in the answer
51	I remembered that we have case studies in class
52	and we discussed about cultural heritage
53	I put in something I have learnt in class if I find it appropriate
54	There might be something from Geography
55	for example, air pollution causes weathering of rocks
56	or how tourists damage cultural heritage
57	seems that I learnt it in class
	(When you went from tourists damaging cultural heritage to global concern, don't you think that the relationship should be explained further?)
58	for the global concern, I'll say that tourists do not understand local cultures well
59	they are not aware of the fact that cultural heritage should not be damaged
60	causing the problem
61	and I will also say that many mainlanders urinate in public places when travelling in Hong Kong
62	this can also be an example
63	it is found that tourists do not know the local etiquettes
64	or they don't know basic etiquettes
	(Don't you find a gap in your explanation from air pollution to global concern?)
65	the carbon dioxide emissions changes immensely from 2012 to 2030…climate change is a problem nowadays
	(You did not mention it in your answer just now)
66	Yes. I forgot
67	as more and more tourists, more and more carbon dioxide
68	it will increase the greenhouse effect
69	the problem is serious now
70	if we continue to do it, it will be more and more serious
	(What will be the problem brought by impolite and disrespectful behaviours?)
71	damaging cultural heritage
72	cultural heritage is valuable and one should not carve his name on it
73	if this is allowed, all other people may do it
74	after hundreds of years, the cultural heritage will disappear
	(you did not elaborate along this line when answering the question, did you?)
75	No
	(When you say the cultural heritage is valuable, are you referring to monetary values?)
76	it is because it is expensive

77	but it is old… passing through hundreds and thousands of years
78	we need to keep it for memory
	(Why didn't you think of this when answering the question?)
79	I think that destruction to cultural heritage is a concern already
	(Don't you think that destroying cultural heritage means destroying a tourist spot for entertainment?)
80	cultural heritage might have a history of several thousand years
81	there is no reason to demolish it
82	you want your next generations to see it
83	we do not know how the Pyramids in Egypt were constructed
84	archeologists need to study it
85	cultural heritage is worth-studying and should not be destroyed
86	the next generations would like to see it
	(Have you got some discussion on cultural heritage and reasons for preservation and cultural values?)
87	I remember that we discussed cultural globalization in class
88	Chinese culture has disappeared as we celebrate Christmas, Halloween and Valentine's Day
89	it may erode the Chinese culture
90	as there are more Starbucks in China, we discussed whether the Chinese culture will vanish
91	but it won't
	(Did you discuss the values of the Chinese culture?)
92	No
	(Besides culture, have you got any discussion on values under other topics?)
93	we discussed whether working female should freeze their eggs
94	and about public health… do you agree with abortion
95	and about family…
	(What about respect for cultures?)
96	No. This is my own view
	(What do you think of your relationship with other people? Mutually beneficial, guarding against the others or competing?)
97	I think I guard against the others
98	I think that resources are limited
99	richer countries can develop renewable energy
100	I think that it is a competition
101	if a country has oil, it will be the richest
102	countries with more advanced renewable energy resources are more prestigious
	(Do you share your notes with the others?)
103	Yes.
104	in reality, it is difficult for countries to cooperate to develop renewable energy resources
105	e.g. China wanted to purchase oil from a certain country but that country refused
	(Why sharing notes?)
106	my teachers said that we are not competing with our classmates
107	we are competing with students outside the school
108	if we share our notes, we may improve together
109	my classmates also share their stuff with us
	(What about your relationship with the environment?)
116	I think it must be in harmony as we are destroying the environment seriously
117	I am afraid that our survival will be threatened
118	like no clean drinking water

Appendix V

Box 1: Email correspondences of consent for the use of secondary sources from the HKAGE

From: Gloria Leung
Sent: Wednesday, March 21, 2018 2:18 PM
To: Dr Fung
Subject: Live Script Study

Dear Dr Fung,

Sorry to bother you about the Live Script Study again.

In order to fulfill the research ethics requirements of the University of Bristol, most grateful if the Research Division of the HKAGE could grant me a written consent for the use of the data from the Live Script Study (including the live scripts, the audio files of the group discussions among the examiners and the think-aloud protocols) in my EDD research.

Million thanks in advance for granting me a consent!

Sorry again for any inconvenience that this may cause.

Regards,
Gloria

From: E. Fung
Sent: Wednesday, March 21, 2018 2:27 PM
To: Gloria Leung
Subject: RE: Live Script Study

Dear Gloria,

With regard to the study captioned above, we are pleased to share the raw data (including the live scripts, the audio files of the group discussions among the examiners and the think-aloud protocols) with you for your research as far as no personal identifications would be revealed when you publish any of your findings.

I wish you success with your studies.

Best regards

E. Fung
Head of Research Division

The Hong Kong Academy for Gifted Education

www.hkage.org.hk

Appendix V

Box 2: Email correspondences of consent for the use of live scripts from the HKEAA

From: C Lee (D-PE)
Sent: Thursday, March 22, 2018 2:22 PM
To: Gloria Leung
Subject: RE: Research study in LS

Dear Gloria,

I'm happy to know about your research study related to HKDSE LS and will support it. Your request for the use of 72 2015 HKDSE LS answer scripts and 72 samples of performance on the HKEAA website is approved in principle but I need a formal letter of support for your research study from your EDD supervisor at the University of Bristol before official approval can be given. Also, permission will be granted by the HKEAA subject to the following conditions

Conditions

a. The permission is granted for educational and non-profit making purpose;

b. For any usage other than the one specified in your request, separate application must be submitted;

c. Permission granted is non-exclusive;

d. Permission does not extend to any third party copyright material (if any) that may be included in the materials;

e. Agreement to share the results of this study with HKEAA for academic use; and

f. Providing HKEAA with a copy of the research findings (and/ or academic publication) upon completion of the study. (Delivery to HKEAA Finance Division – Support Services, Room 1311, 13/F Southorn Centre, 130 Hennessy Road, Wan Chai, HK)

The HKEAA is not in a position to grant approval for the use of the transcripts of group discussions among examiners as we do not hold the copyright of the transcripts. I think you need to seek the consent of the participants of the group discussion before recording their discussion and also their approval for you to use the transcript for your study. I would also like to ensure that their participation in the group discussion will not compromise the confidentiality of their appointment as HKDSE examiners.

Thanks and regards,

C. Lee

Appendix V

Box 3: Email correspondences of consents from the examiners for the use of the scores and the transcripts of nominal group discussions

On Wed, Mar 21, 2018 at 10:22 AM, Gloria Leung wrote:

Dear XXX,

Million thanks for your time and professional input in the live script study in 2016!

As mentioned during the meetings, the data collected will be used for research purposes by both the HKAGE and myself. I am now entering the thesis phase of my EDD study. My research topic is "An Evaluation of the Validity of a Large-scale Assessment". To fulfill the research ethics requirements, most grateful if you could grant me a written consent for the use of your scores on the live scripts and the transcripts of the group discussions for my EDD research. Hope to share with you my findings on the LS Examination in due course.

Looking forward to your email granting me your consent!

Million thanks indeed!

Regards,

Gloria

From: Examiner 1
Sent: Wednesday, March 21, 2018 11:37 AM
To: Gloria Leung
Subject: Re: Live Script Study with HKAGE

Dear Gloria,

It's my pleasure and agree to share the scores on the life scripts and discussion transcripts for the research purposes.

All the best!

Best regards,

Examiner 1

To whom it may concern,

I, Examiner 2, give my consent for the research topic of 'An Evaluation of the Validity of a Large-scale Assessment' to use of my scores on the live scripts and the transcripts of the group discussions for Ms. Gloria Leung's EDD research. Any question regarding this consent can be directed to me using the following phone no XXX.

Best regards,

Examiner 2

Box 4: Email correspondences of consents from the examiners for the use of the scores and the transcripts of nominal group discussions (continued)

From: Examiner 3
Sent: Wednesday, March 21, 2018 8:54 PM
To: Gloria Leung
Subject: 回覆： Live Script Study with HKAGE

Dear Gloria,

I agree to share my scores on the live scripts and the transcripts of the group discussion for research purposes.

Wish you all the best with the research.

Best regards,

Examiner 3

From: Examiner 4
Sent: Wednesday, March 21, 2018 3:33 PM
To: Gloria Leung

Subject: Re: Live Script Study with HKAGE

Dear Ms Leung Tsz Yim Gloria,

I agree that my scores on the live scripts and the transcripts of the group discussions can be used for research purpose.

All the best!

Best regards,

Examiner 4

Box 5: Consent Form for the participants of the think-aloud study

Think-aloud Study

Consent Form

Participant's Name: XXX

- **I understand that the data will be used anonymously.**
- **I authorise the researchers of this study to make use of the data contributed by me for research purpose.**

Student's signature: XXX Date: 22 August 2015